DOCUMENT RESUME

ED 191 920 TM 800 546

AUTHOR	Johnson, Sylvia T.
TITLE	The Measurement Mystique: Issues in Selections for Professional Schools and Employment. Occasional Paper, 1979, No. 02.
INSTITUTION	Howard Univ., Washington, D.C. Inst. for the Study of Educational Policy.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	79
NOTE	70p.
AVAILABLE FROM	Institute for the Study of Educational Policy, Howard University, 2935 Upton St. N.W., Washington, DC 20008 (Free).
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Achievement Tests; *Admission Criteria; Aptitude Tests; Blacks; Females; Higher Education; Measurement; *Medical Schools; *Minority Groups; *Professional Education; *Test Bias; *Testing Problems; Test Reliability; Test Validity; Whites
IDENTIFIERS	Bakke v Regents of University of California

ABSTRACT

This paper grew out of a presentation made as part of a panel discussion on the Bakke case, and is second in a series of Institute for the Study of Educational Policy documents on important educational issues and problems. Many of the factors that impact test performance or scores are examined. The primary thesis of the paper is that it is psychometrically legitimate and practically sound to supplement test scores when tests are used as selection devices in a decision-making process. Limiting dependence on tests is a practice that will result in greater fairness to more individuals. The Bakke case is used as the source of basic data. A discussion of the factors affecting test scores is presented, and the issues of validity and selection are discussed. Specific areas of concern are: measurement for selection; origins and identification of test bias; bias resulting from improper use; using regression analysis in determining predictive validity; and the issue of overprediction. Recommendations supporting the paper's thesis are offered. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

APR 15 1980

ED191920

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE MEASUREMENT MYSTIQUE

SYLVIA T. JOHNSON
Senior Visiting Fellow
ISEP
Associate Professor,
School of Education
Howard University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
K. Tollett

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800 546

OCCASIONAL PAPER
of the
Institute for the Study of Educational Policy

1979, No. 02

Howard University
Dunbarton

THE "MEASUREMENT MYSTIQUE" - ISSUES IN SELECTION FOR PROFESSIONAL SCHOOLS AND EMPLOYMENT

Sylvia T. Johnson
Senior Visiting Fellow
ISEP
Associate Professor
School of Education
Howard University

This publication was made possible by a grant from the
Ford Foundation

COPYRIGHT
INSTITUTE FOR THE STUDY OF EDUCATIONAL POLICY
1979

To my parents,
Katie Taylor Moore and
Henry Starns Taylor
whose effects are immeasurable.

TABLE OF CONTENTS

Page

List of Tables vi
List of Figures vi
Foreword vii
Acknowledgements xi
I. Measurement for Selection—
An Overview 1
A. Introduction 1
B. The Nature of Measurement 2
C. The *Bakke* Case 3
D. Research on Medical School Admissions 6
E. The Search for a Criterion 10
F. Use of Tests in Selection 12
II. Test Bias - Origins and Identification 15
A. Psychometric Factors in Test Bias 17
1. Item Content as a Source of Bias 18
2. Subject Characteristics as Factors in Test Bias 20
a. Race, Racism, and Test Bias 20
b. Motivation as a Factor in Test Bias 21
c. Socioeconomic Status 27
3. Setting and Administration Factors in Test Bias 28
a. The Tester 28
b. Physical Surroundings and Bias 29
B. Bias Resulting From Improper Use 29
1. Bias in the Selection Rule 31
2. Bias in the Meaning Given to a Test Score 33
C. Validity and Test Bias 34
D. Reliability and Test Bias 36
III. Using Regression Analysis in Determining Predictive Validity 38
IV. The Overprediction Issue 44
V. Conclusions and Recommendations 46
Bibliography 49

LIST OF TABLES

Table — Page

1. Expectancy Table for Medical School Progress Based on Verbal Ability MCAT Scores of 1949 - 1958 Entrants 8
2. Expectancy Table for Medical School Progress Based on Science Achievement MCAT Scores of 1949 - 1958 Entrants 8
3. Medical College Admission Test Scores, Undergraduate Grade-Point Averages, and Acceptance Rates of Applicants by Parental Income, 1975 - 1976 First Year Class 27
4. Mean MCAT Scores of Accepted and Rejected Applicants 37

LIST OF FIGURES

Figure — Page

1. Sources of Test Bias 16
2. Predicted and Actual Performance 39
3. Actual Performance for Persons Earning Same Test Score 40
4. Predicted Performance Levels of Selected and Non-Selected Applicants 41
5. Prediction Diagram I 42
6. Prediction Diagram II 43
7. Prediction Diagram III 43

FOREWORD

As the decision in the *Regents of the University of California v. Bakke* approached, the National Advisory Board of the Institute recognized that an underlying problem in *special minority admissions* programs was the belief, if not the conviction, that there was something fundamentally unfair in admitting minority students who had lower test scores than majority/white students. With almost no exception, the members of the Board were familiar with the consideration of many attributes, characteristics, and factors other than standardized test scores in the admissions process throughout higher education, be it undergraduate, graduate, or professional. Thus, the Board advised the staff to review the state of the art in testing and measurement as they relate to the admissions process. In compliance with this directive, Dr. Sylvia T. Johnson, an expert in this field, was retained as a Visiting Fellow in the summer of 1978 to look at tests in the selection process for professional schools.

The product of Dr. Johnson's work, entitled *The "Measurement Mystique": Issues in Selection for Professional Schools and Employment,* follows and becomes the second in a series of ISEP occasional papers and research documents on important educational issues and problems. The first occasional paper published in September of 1977 entitled *Higher Education's Responsibility for Advancing Equality of Opportunity and Justice* by Dr. James E. Cheek, President of Howard University, launched the series. Dr. Johnson's paper follows in the tradition of excellence of Dr. Cheek's, however, it moves through the labyrinth of psychometry with its concern for validity and reliability, dissecting regression and other powerful logical tools of analysis. Nevertheless, the grand policy and value analysis of Dr. Cheek's paper is not irrelevant to Dr. Johnson's paper.

Of the many important and striking observations made by Dr. Johnson in her paper, it is very noteworthy that she states, "The choice of a criterion of interest is . . . fundamentally a

policy decision, and not a psychometric one." Thus, are we interested in instruments which will predict the first or second year performance in medical or law school or do we want instruments which will predict what kind of doctor or lawyer an applicant will become, showing how much interest in and commitment to a particular clientele? Furthermore, the gross under-representation of Blacks, Chicanos, and females in the professions is not irrelevant to the admissions process. Therefore, as Dr. Johnson writes, "If decision makers feel that it is useful and important to have fair numbers of Black and female physicians, then a utility function would give weight to the minority or sex status of applicants."

Dr. Johnson canvasses most of the factors that impact test performance or scores. Of great interest to those concerned with improving educational opportunity is her observation that Black children perform significantly better under warm testing situations. She states, "This means that warmth (and possibly other related interpersonal characteristics) had its greatest effect on the performance of low SES Black youngsters." The implications of this are far-reaching.

The staff and Advisory Board deeply appreciate Dr. Johnson's advancement of our understanding of the *measurement mystique.* As a research and policy center, ISEP has three program objectives:

- To prepare a periodic critical assessment of the dynamic status and needs of Blacks in higher education
- To assess the impact of law and social science and other research on the status of Blacks in higher education
- To use old models creatively and to develop new models and paradigms for the explication of higher education and implications for elementary and secondary education.

Through its monographs and annual reports, through its seminars and conferences, as well as through its occasional publications and public testimony, the Institute attempts to fill

a vacuum in the organized body of knowledge about higher educational opportunity for Blacks and other minorities. In so doing, ISEP attempts to make a significant contribution to the formulation and evaluation of contemporary educational policy.

Howard University
Washington, D.C.
February, 1979

—Kenneth S. Tollett, Chairman
National Advisory Board,
Institute for the Study of
Educational Policy

NATIONAL ADVISORY BOARD

Kenneth S. Tollett, Chairman
Distinguished Professor of Higher Education
Howard University-Dunbarton Campus
Washington, D.C.

Elias Blake, Jr., President
Clark College
Atlanta, Georgia

James E. Cheek, President
Howard University
Washington, D.C.

Lewis C. Dowdy, President
North Carolina A & T University
Greensboro, North Carolina

Edward Hawthorne, Dean
The Graduate School
Howard University
Washington, D.C.

Willis T. Howard, Dean
School of Education
Howard University
Washington, D.C.

Clark Kerr, Chairman
Carnegie Council on Policy Studies in Higher Education
Berkeley, California

James R. Lawson, Director
Office of University Affairs
National Aeronautics and Space Administration
Washington, D.C.

Lionel Newsom, President
Central State University
Wilberforce, Ohio

James Perkins, Chairman
Board of International Council for Educational Development
New York, New York

Lois Rice, Vice President
College Entrance Examination Board
Washington, D.C.

Norvel Smith, Associate Vice Chancellor for Student Affairs
University of California-Berkeley

Nancy H. Teeters
Member of Board of Governors
Federal Reserve System
Washington, D.C.

Horace Traylor, Dean of Development
Miami-Dade Community College
Miami, Florida

Lorraine A. Williams, Vice President for Academic Affairs
Howard University
Washington, D.C.

Stephen Wright, Senior Advisor to the President
College Entrance Examination Board
New York, New York

ACKNOWLEDGEMENTS

This paper grew out of a presentation made as part of a panel seminar on the *Bakke* case, sponsored by the Graduate Student Council of the School of Education at Howard University. Among the members on the panel were Professors John E. Fleming and Faustine Jones. Dr. Fleming is a Senior Research Associate at ISEP and his suggestions as well as feedback from the ISEP research staff were instrumental in the development of this paper. Dr. Jones' comprehensive text, *The Changing Mood in America: Eroding Commitment?*, published by Howard University Press, also provided strong impetus for ideas presented here.

Special thanks must go to ISEP Director, Kenneth S. Tollett, for providing the opportunity to spend a summer at ISEP as part of the research staff. Not only did this assist in developing the paper but also it provided a unique opportunity to see the research potential of the ISEP organization.

In the development and expansion of this paper from a brief panel statement to its present form, the suggestions of my husband, Dr. Howard N. Johnson, helped to define the major areas to be developed. For this help, as well as continual less documentable assistance, I am thankful.

Earlier drafts of the manuscript were read by Professors Howard Cameron and Charles Asbury of the School of Education at Howard University, Professor William E. Coffman of the University of Iowa, and Dr. Davis G. Johnson of the Association of American Medical Colleges. Their helpful comments and suggestions were appreciated and largely incorporated. Also, the permission of Dr. Johnson and the AAMC for the use of tables 1-4 is gratefully acknowledged.

The supportive and thorough editorial assistance of Earnestine Stripling, new to ISEP but rapidly assimilated, was

essential in achieving the final product.

The opportunity to get feedback from the impressive ISEP National Advisory Board was also a helpful experience and several relevant sources for the text were located through Board members. The assistance of Dr. Norvel Smith in locating a Berkeley Task Force Report was especially useful.

The Dean of the School of Education, Dr. Willie T. Howard and my Department Chairperson, Dr. Mauderie H. Saunders, have also been supportive of my efforts in completing this project.

General clerical supervision was ably furnished by Mrs. Margaret Nealy and much of the typing was completed by Ms. Harrietta Fowler, Ms. Sandra Davis, and Mrs. Gloria Johnson.

The support of the Ford Foundation in this endeavor is acknowledged and appreciated for without its support, this paper would not have been possible. However, the ideas expressed in this paper do not necessarily reflect the position of the Ford Foundation.

I. MEASUREMENT FOR SELECTION - AN OVERVIEW

A. INTRODUCTION

Measuring characteristics of people has a certain aura or mystique about it. The idea of moving from a formula to a number, then to the meaning of that number represents to some a psychological sleight-of-hand to be admired but never fully understood or trusted.

Meanings conveyed by statistics have received limited study by statisticians. Kruskal (1978) noted:

> It seems ludicrous for us (statisticians) to be concerned about exact significance levels, accuracy of standard errors, robustness of Bayesian procedures against fluctuation in prior beliefs, . . . and yet to throw away all that exactness through our lack of concern about what gets communicated in the words we use—to each other and to our wider publics. (p. 229)

It does seem that tests and testers arouse strong feelings in many people, and that these feelings relate to what the tests mean. It is not simply that some are in favor of tests and some against them, although these groups do exist. It is more common to find many people with strong ambivalent feelings. Statements such as "tests are unfair," "tests label kids," "tests invade privacy," "tests have little meaning" can draw applause. The very same audience will then argue for a strong meaning of a score; asserting that if one person's test

---

"Perhaps tests need not be labeled like cigarettes with a health hazard warning in bold face type on the cover, but certainly, the limitations on their usage should be noted."

---

score is higher than that of another, the first should be favored whatever the reward or benefit may be. Such a meaning supposes that the test score is an implicit substitution for the "real thing" that we want to measure. The problem clearly shows the need for carefully examining the issues related to tests, test usage, and the meaning that come from them, and informing the public in a meaningful way. Perhaps tests need not be labeled like cigarettes with a health hazard warning in bold face type on the cover, but certainly, the limitations on their usage should be noted.

The *Bakke* case has pointed up the need for greater understanding of what educational measurement can and cannot do. The court decision could lead to inappropriate changes in admissions policies in professional schools, if it is seen as negating some sound practices of selection which employ test and non-test data in decision making. The proper role of measurement professionals and the testing instruments which they have developed in making decisions that affect the lives of other people should be examined in light of this case.

The primary thesis of this paper is the following:

> It is psychometrically legitimate and practically sound to supplement the test score, when a test is used as a selection device, and in a decision-making process. There is adequate precedent for this practice, there is theoretical support, and there is widespread current use of broad-based selection systems. Limiting dependence on tests is a practice that will result in greater fairness to more individuals, as well as groups, in the long run.

B. THE NATURE OF MEASUREMENT

Since we cannot measure minds in the same way that we measure the length of a room, all educational and psychological measurement is done indirectly. By measuring performance or behavior on some task, we infer the level of achievement, intellectual functioning, or some other psychological construct. But because measurement is indirect does not mean it is faulty. In fact, much of physical measurement is done indirectly, and inferences are made because of certain theories about how the universe operates. These inferences make it possible to measure everything from distances in the solar system to distances between and within atoms.

Physical measurement as well as psychological measurement can result in errors. However, our errors in psychological measurement may be particularly hard to identify and correct. If I theorize that the combination of ingredients used in a cement wall will not harden, and I attempt to test this theory by walking through this area, I will meet with solid evidence to refute my theory. The social scientist lacks the benefit of such firm validation (or invalidation), and is faced with a stronger responsibility for careful, thoughtful analysis and empirical validation. In

testing, the behavior measure used as the basis for an inference is the test score. An inference is then made to a level of achievement, aptitude, personality, interest or some other human characteristic which cannot be measured directly.

The relationship between the test score and the characteristic which we wish to infer may not be the same for all examinees, particularly for many Blacks and other racial minorities who have been systematically excluded from educational benefits. As a result, test scores alone frequently are not sufficient measures of aptitude or other psychological constructs for minority groups. Leaders in minority communities have seen numerous cases of discrepancy between actual ability to perform and measured aptitude, and many regard tests as useless. The reality is that tests are increasingly used in all forms of decision making, and that their uses and misuses need to be more fully understood by policy makers and community leaders.

Many factors operate to attenuate or lower test scores, and these factors tend to have their greatest effects on Blacks and other minority applicants. These include factors which affect the actual performance of individuals on the test, such as socioeconomic status, differences in educational opportunity, motivation, narrowness of content of the tests, atmosphere of the testing situation, and the perceived relevance of the test to success. They also include factors that affect the test score more indirectly, such as the composition of the group used for item tryouts and item selection and analysis which precede the actual standardization, composition of the standardization or normative group, and the techniques and procedures employed in item construction. Also, the validity or appropriateness of tests often differ for Black and white applicants, in relation to the same future performance or criterion. These attenuating factors and the resultant validity problem encountered in measuring scholastic aptitude will be discussed in this paper.

C. THE *BAKKE* CASE

In exploring this position, the ***Bakke*** case will be used as the source of basic data, since there is a large volume of research on medical school selection, and the Supreme Court decision has raised the awareness of the public to this case. This paper will present a discussion of the factors affecting test scores. Then, the issues of validity and selection will be discussed.

The case of *The University of California Board of Regents v. Allan Bakke* has received extensive coverage and been researched from differing viewpoints. The decision of the Supreme Court hinged on the dual nature of the admissions procedure at the Davis campus. In its narrow ruling, the Court ruled that race could be considered in admissions, but not by the procedure of two separate admissions routes which the University of California had employed. There is a danger here that the decision may be misunderstood, and the wrong inferences drawn.

The issue which concerns people in general is a broader one. Should a good test score earn one admission, and does it mean that one is better qualified than those with lower scores? The crux of the case in the "public mind" seems to hinge on the meaning of the benchmark score, which was a composite of several predictors. The major component of this benchmark score was the Medical College Admission Test (MCAT), a psychological instrument developed for the Association of American Medical Colleges (AAMC) by the American Institutes for Research (AIR), and administered by the American College Testing Program (ACT).

As would any applicant to medical schools in the country, Allan Bakke took the MCAT. He then requested that his score be sent to the University of California at Davis, as well as to eleven other medical schools to which he had applied. Bakke was rejected by all twelve schools, and twice rejected by the University of California at Davis. He filed suit against Davis, claiming that a special admissions program operated to exclude him on the basis of race in violation of the Equal Protection Clause of the 14th Amendment. The Supreme Court of the State of California held that the special admissions program at Davis violated the Equal Protection Clause. The United States Supreme Court ordered Bakke's admission, and invalidated that aspect of the special admissions program that reserved a specific number of spaces for disadvantaged minorities.

A series of closely related questions can be examined in terms of the role of the test in the *Bakke* admissions case:

1. Is the claim that Bakke's Medical College Admission Test Score "entitled" him to admission to medical school valid, from consideration of the test itself?

2. To what extent does psychometric theory support admitting to medical school only those persons having the highest test scores?

3. To what extent is the use of alternate criteria for admission, (that is, criteria other than tests scores and grade point averages) supported by psychometric theory?

Recognizing the history of adverse conditions which Blacks have had to endure as they sought higher education, the case for the use of alternate criteria for selection in medical school admission has been made on multiple grounds. Foremost of these is the fact that there is a societal need in our country for more Black and other minority physicians. If this need were met, the poor ratio of residents to physicians in low-income minority communities, particularly in urban and rural America, could be decreased. Clearly, the providing of medical services to minority communities is the responsibility of all practitioners, but the data supports the fact that most medical practitioners serving minority communities are from minority groups, and that minority physicians serve mainly minority patients (Montoya, 1977; Abarbanel, 1976; Lloyd et al. 1978). A concomitant outcome of an increase in physicians in these communities would be the provision of additional role models for Black youth. Also, weaker academic preparatory resources, particularly in the sciences, have been available to Blacks as a function of lower funding levels for schools serving Blacks, and lack of effective community pressure. Further, there is still extra-academic stress encountered in simply going to school (Willie et al., 1973; Barbarin, 1977). The latter point was eloquently made by a young University of California at Davis Medical School honor graduate who completed her undergraduate work at Stanford in what she described as an atmosphere characterized by slurs and other outrages (Kotkin, 1978).

Yet, little has been said about the fundamental weakness, from the standpoint of educational and psychological measurement, of using only test scores as a criterion for the selection of students. Equally little has been said about the benefits of incorporating other criteria into the selection process, again from a psychometric point of view. The developers of the MCAT and other selection tests do note clearly the limitations of their instruments, but the general public may not be aware of these limitations (Cuca, 1976). Often in the public mind, the test score is ability itself.

D. RESEARCH ON MEDICAL SCHOOL ADMISSIONS

A major producer of research on medical school admissions is the Association of American Medical Colleges (AAMC). Through their Divisions of Student Studies, and of Educational Measurement and Research, as well as their fifty-year-old publication, *Journal of Medical Education*, this organization has encouraged, conducted and disseminated such research. In examining medical school admissions patterns during the last 40 years, their studies have documented the relationships between the medical school admissions process and the meeting of institutional and national objectives, and identified a wide range of predictors used in selection (Cuca et al., 1976).

During World War II, the national need for military physicians resulted in an 80% military, and 20% nonmilitary quota system for medical schools. According to Cooper as cited in Cuca et al. (1976), the procedure used to meet this pressing national need was the following:

> . . . The selection of medical students was made primarily by Army and Navy officials in consultation with medical school admissions officers. Since the premedical program was also largely under the auspices of the military, it was possible to evaluate students within each region and to assign them to a medical school within that region. During the War, 80 percent of the available places were filled by this admissions technique. The remaining 20 percent of the students were selected solely by the medical schools from draft-exempt men and from women. (p. 1)

Other research (Cuca et al, 1976) within the medical education community has examined the utility of a wide range of factors in medical school selection, and confirms that most medical schools actually employ a range of predictors in selecting students. For example, the Florida State Program in Medical Sciences (Elliot, 1975) employs a selection model which gives 44% of the total weight to academic factors, and 56% to nonacademic factors. The factors involved include grade point averages, MCAT scores, faculty recommendations, socioeconomic and demographic ratings, Myers-Briggs Type Indicator Sensing Scales, academic distinctions, and interviews.

The conclusion that the academic performance and test scores may have different meanings for different individuals, depending on other factors, is suggested by results of a study by Schofield and Merwin (1966), and reviewed by Cuca et al., (1976). Their procedure was to examine predictors which had the greatest association with criteria, and then to determine regression weights which would optimize their prediction of criteria. They found that in the setting examined, the pre-medical performance of applicants should be given different weighting, depending on the direction of interest and personality variables. Thus, if favorable interest and personality measure results were found, lower weights of academic variables resulted in better predictions. On the other hand, if deviant interest and personality scores were achieved, the academic variables should be given higher weightings.

...e 1976 review by Cuca et al. discusses extensively the use of noncognitive criteria for selection. It notes that the use of noncognitive criteria is currently growing in the selection of minority students, but not for non-minority students. The researchers conclude that noncognitive criteria should be given greater consideration in selection of non-minority applicants, since the similarity of these applicants on the typically measured cognitive predictors is so great.

The similarity between test scores of "doctors and dropouts" is shown in a comprehensive attrition study by D. Johnson and E. Hutchins (1966), who analyzed data for all member schools of the Association of American Medical Colleges for the period 1949-1958. The study showed that although there was a relationship between test score and performance when the entire range of scores was considered, the chance of success for a student with a middle range verbal ability or science achievement score was only slightly lower than the chance of success for a student scoring 150-200 points higher. For example, the chances of success in completing medical school of a student with a verbal ability score in the range of 450-499 were 91 out of 100, while the corresponding chances of success for students scoring in the 650-699 range were 93 out of 100 (Table 1). In the expectancy table for the Science Achievement MCAT, the differences were somewhat larger. The chances of success for students scoring in the 450-499 range were 90 out of 100, while students scoring 650-699 had 95 out of 100 chances for success in achieving an M.D. degree (Table 2). In addition, a report of the American Medical Colleges (Johnson, Smith, Tarnoff, 1975) shows little difference in scores for successful and non-successful admitted students in 1970. The data also show that there is little

difference between the MCAT scores of Caucasian males who were promoted, and those who were not.

Table 1

EXPECTANCY TABLE FOR MEDICAL SCHOOL PROGRESS BASED ON VERBAL ABILITY MCAT SCORES OF 1949-1958 ENTRANTS

		Chances in 100 That Student Will Make Progress Indicated						
If VA Score Is Between:	Based on N of:	Total M.D.	On Time M.D.	Delayed M.D.	Repeat M.D.	Academic Dropout	Nonacademic Dropout	Total Dropout
750-799	197	94	88	2	4	2	4	6
700-749	1,465	94	88	3	3	3	3	6
650-699	5,192	93	87	3	3	3	4	7
600-649	8,868	93	87	2	4	3	4	7
550-599	12,151	93	87	2	4	4	3	7
500-549	13,895	92	87	1	4	5	3	8
450-499	13,328	91	85	1	5	5	4	9
400-449	9,604	90	84	1	5	7	3	10
350-399	4,558	87	79	1	7	9	4	13
300-349	1,361	82	72	1	9	13	5	18
250-299	206	80	69	2	9	16	4	20
200-249	15	87	74	0	13	13	0	13
No MCAT	4,613	88	82	1	5	6	6	12
Total	75,453	91	85	1	5	5	4	9

Table 2

EXPECTANCY TABLE FOR MEDICAL SCHOOL PROGRESS BASED ON SCIENCE ACHIEVEMENT MCAT SCORES OF 1949-1958 ENTRANTS

		Chances in 100 That Student Will Make Progress Indicated						
If Science Score Is Between:	Based on N of:	Total M.D.	On Time M.D.	Delayed M.D.	Repeat M.D.	Academic Dropout	Nonacademic Dropout	Total Dropout
750-799	396	96	91	3	1	1	4	4
700-749	1,731	96	90	3	3	1	3	4
650-699	4,539	95	90	3	2	2	3	5
600-649	8,859	95	90	2	3	2	3	5
550-599	12,586	94	90	1	3	3	3	6
500-549	15,135	93	88	1	4	4	3	7
450-499	13,202	90	84	1	5	6	4	10
400-499	9,057	87	80	1	6	9	4	13
350-399	3,973	82	73	1	8	12	6	18
300-349	1,158	75	63	1	11	19	6	25
250-299	167	63	52	2	9	26	11	37
200-249	31	68	68	0	0	19	13	32
No MCAT	4,619	88	82	1	5	6	6	12
Total	75,453	91	85	1	4	5	4	9

Note. From "Doctor or Dropout? A Study of Medical Student Attrition," by D.G. Johnson and E.B. Hutchins, *Journal of Medical Education*, 1966, *41*, 1097-1269. Reprinted by permission.

The relationship between MCAT scores and promotional status for Blacks and women during 1970 differ from those of the majority. The results for Blacks show greater differences in MCAT scores between promoted and non-promoted students than do the scores for white males. These results suggest two possible conclusions. First, there may be differential validity for the MCAT according to race; secondly, the MCAT may have good validity for predicting promotion at the end of the freshman year for scores in a certain range, and limited validity in other parts of the range, for both Blacks and whites.

The results for white females also bear close study. The MCAT scores for white females who were not promoted were greater than the scores of those who were. It would seem that factors negatively related to the MCAT affected their promotion to the second year medical school class, at least in the year 1970. The results of a single year cannot be taken as an index of overall validity, but if these results were to continue, they would indicate low validity for the use of the MCAT for white females.

The Association of American Medical Colleges has recently revised its admissions test radically. The New MCAT differs from the "old" MCAT taken by Bakke and all other applicants prior to 1977 in content, scaling, and potential usage. The New MCAT may be a better instrument, but the basic principles of selection and prediction outlined here still apply. The New MCAT was designed to measure problem solving and scientific reasoning skills, as well as certain prerequisite knowledge. According to the interpretive manual (AAMC, 1977), the goals employed in its development were as follows:

1. The development of an assessment instrument to measure achievement levels and the expected prerequisites which are generally relevant to the study and practice of medicine.

2. The design of an instrument which would assess a precisely identified base of content manifested in all test forms and also possess the capability to respond to a changing base of scientific knowledge.

3. The construction of a test whose results would identify strengths and weaknesses and thus aid in the counseling of examinees.

These goals imply a multi-dimensional instrument based on specific subject matter content areas. However, a factor analytic study of the New MCAT, based on a national sample of 30,599 examinees, showed very high loadings of all subtests on a single factor, accounting for 71% of the total common variance (Molidor et al., 1978). The intercorrelation matrix of the subscales was used in the factoring. Similar work with the matrix of correlations between the items, rather than between subscales, should be useful in trying to determine the factors underlying the instrument. If the Molidor results are supported by subsequent research, the New MCAT could be less useful than the "old" MCAT in meeting the goals of its development.

E. THE SEARCH FOR A CRITERION

The choice of a criterion of interest is perhaps the most difficult and important step in the entire validation process. It is fundamentally a policy decision, and not a psychometric one. The question is, what is that important thing in the future that we want our selection instruments to predict? Is it "ability to survive the first year of medical school?" Is it "lifelong involvement in pure research?" Or is it "personal ego strength and emotional stability sufficient to be involved as a psychotherapist for 40 years?"

"The choice of a criterion of interest....is fundamentally a policy decision, and not a psychometric one."

The choice of a criterion can be considered as having different stages (Astin, 1964). First, a conceptual criterion must be developed. The conceptual criterion is a verbal statement of outcomes that are important or socially relevant. It is based on the general purposes of the sponsor of the research, which may be a professional organization, a firm, or the researcher herself.

After a conceptual criterion is identified, the task is to determine a criterion performance. A criterion performance is any observable event which is judged relevant to the

conceptual criterion. From these performances, the criterion measures are derived. These may be indices of clinical performances, test scores, number of job changes, employer or patient ratings, etc. The criterion measure is an attempt to operationally define, insofar as possible, the conceptual criterion. It is always important to focus on the actual conceptual criterion, even though it is just a statement, and not the operational definition. At some point, we may be able to reach closer to the conceptual criterion choosing other criterion measures, or improving existing ones. We do not validate criterion measures or criteria. We validate predictors. A criterion measure can only be validated by logical analysis of its relevance to the conceptual criterion (Astin, 1964).

"Maintaining the conceptual criterion focus is akin to 'taking a long look' on the highway, rather than following the car immediately in front of you. It is easy to look at your current criterion measure, such as freshman year success as the thing you are actually trying to predict, and lose sight of the long term goal of predicting performance as a physician."

Maintaining the conceptual criterion focus is akin to "taking a long look" on the highway, rather than following the car immediately in front of you. It is easy to look at your current criterion measure, such as freshman year success as the thing you are actually trying to predict, and lose sight of the long term goal of predicting performance as a physician.

The definition of conceptual criteria for professional school selection is not easy. Do we always want to produce the same type of physician? Involvement and interests by medical school students in recent years in family practice and community medicine would appear to be a trend of positive benefit to the medical care of the average citizen. At the same time, physicians are needed to teach in medical schools and be involved in the "cutting edge" of medical research laboratories and centers. One type of physician is not "better," in terms of their utility to community, but each of these aspects of physician performance impact on health care in an important way.

Medical School educators have long recognized the criterion problem, and continue to direct their efforts toward dealing with it. According to Cuca et al. (1976):

> While admissions continues to be a problem of prediction, the criteria being predicted have changed during the last two decades. In the 1950s and early

1960s the major concern of admissions committees was to select applicants who would successfully complete their medical training and go on to be "good" physicians. The concern was to reduce the higher rates of attrition from medical school which were characteristic of the time." (p. 43)

While the attempt to keep attrition down continues to be an important element in the selection process, a greatly enlarged applicant pool from which to select potential medical students, the improved academic credentials of today's applicants, and pressures for the production of physicians to fill societal needs have combined to deemphasize medical school achievement as the major criterion being predicted by the admissions process. Furthermore, there has always existed an awareness in the medical community that there is no one-to-one relationship between medical school achievement and performance as a physician.

One of the major changes evidenced in the literature concerning selection variables over the past twenty years has been the change in emphasis on attempting to predict the immediate objective of medical school achievement to predicting the long-term objective of physician performance, or quality of actual medical practice. A review of research on the measurement of physician performance by Barro (1973) points up the variety of techniques for assessing two dimensions of performance: 1) the process of patient management, and 2) the outcome, or the evaluation of end results in terms of health and satisfaction. The process techniques reviewed include technical, interpersonal, and simulated process approaches. McGuire (1976) developed a system for simulation as an assessment technique in a work that should add impetus to this area. Yet, the question of which of these criteria should serve as a base for the admissions process remains.

F. USE OF TESTS IN SELECTION

A test is a sample of behavior collected under controlled conditions for a specific purpose. When used for selection, tests are useful if they improve the percentage of correct choices over some other selection procedure.

Several specialized aptitude tests are widely employed for selection for graduate and professional schools in the United States. The oldest and most widely used include the

Medical College Admissions Test (MCAT), the Law School Aptitude Test (LSAT), and the Graduate Record Examination (GRE).

The Medical College Admissions Test is administered for use in selecting persons for admission to medical school. Its aim is to assist in selecting those who will be successful in medical school, and rejecting those who would be unsuccessful. There is evidence that it does that to some extent, since students who do not complete medical school on the average have lower MCAT scores than those who succeed.

The problem arises when it is assumed that if a certain cutting score on the test eliminates potentially unsuccessful applicants, choosing a higher cutting score will be "better," thus automatically eliminating more of the potentially unsuccessful. When in fact, as that cutting score is raised, many potentially successful applicants are also eliminated. Given a group of qualified applicants, the fact that one student's test score or grades are somewhat higher than those of another does not mean that the first is more "qualified" than the second. Some basic level of competency should be established, but beyond that many other factors should be used to determine success. A rough analogy can be seen in the following situation: Suppose one wishes to board a city bus to travel some distance in a large city and there is an established bus fare. If you have less than the established fare, you will not get a ride. The exact fare is required. Having more money will not take you farther (under a fixed rate schedule). Whether you reach your destination does not relate to whether you have more than the basic fare in your pocket. It relates to other actions you may take and characteristics which you may have, such as how well you check the route map and bus schedule, and whether you watch to see when you are nearing your destination so that you can head for the exit. Your timely arrival also relates to external factors, such as the traffic situation, possible accidents, etc.

Let us carry our analogy one step further. Imagine an entire collection of individuals are waiting to board this bus. Some have the required bus fare and others do not. Then, the bus fare will be one method of determining who rides the bus. In this situation, there are a whole set of antecedent questions that determine who even gets a chance to drop the fare in the box. For instance, how widely were the bus schedules and route information distributed? How many people can the bus carry? Who regularly rides this route or knows other people

who have taken this bus? Who gets in line first? Who is driving the bus? Once the bus is boarded, the same characteristics and external factors determine success in negotiating the system.

Thus, we can see that a whole set of factors determine success. A good predictor, (in this case the bus fare), can measure some of these factors, but not all of them, and more of one single factor does not necessarily mean that one's chances of success are proportionally increased.

An educational example would be helpful here, but our image of the test as a "mental yardstick" makes acceptance difficult. Let us conceive of the test as a yardstick that varies in appearance, depending on the purpose for which it is used. A medical school aptitude test used to predict the criterion of success as a physician could be viewed as a yardstick with numbers clearly printed in even intervals nearly halfway up the stick, and then blank the rest of the way. The implication? That this is a reasonable measure of what we're interested in, up to a certain point. After that, we need to look at other "yardsticks."

The same test might be a "blank yardstick" when used for another purpose, such as predicting success as a sculptor or musician. It might be a "rubber yardstick" if used to predict success in high school science courses. Even though the type of content is appropriate, the inappropriate difficulty level would result in measurement with low reliability and validity.

The personal attributes and characteristics that are the focus of many tests are complex entities. Test developers and measurement researchers are aware of the multidimensional nature of intelligence and other aptitudes. Horst (1968) noted, "The general impression that intelligence is unidimensional is borne out by the common usage of the word. We speak constantly of individuals as having high or low intelligence, or of having some particular degree of intelligence. This practice is part of our folklore and an example of our tendency to oversimplify complex concepts" (p. 429). Verbal tests with academic type tasks are normally used to measure academic aptitude, general aptitude, and intelligence. The justification of this procedure is that when non-verbal and performance tasks are used, they normally correlate quite highly with verbal tests, and thus the more reliable, easier to administer, and

cheaper verbal tasks are employed. However, the correlations between verbal and performance tasks are not equally high in all groups. They are most likely to be high in those persons who have had the greatest opportunity to develop their verbal skills maximally, those with the greatest environmental resources for the development of traditional verbal skills. The fact that other tasks may more accurately assess aptitude in minority youth is not considered.

II. TEST BIAS-ORIGINS AND IDENTIFICATION

Fair testing has long been considered by psychologists (Anastasi, 1950; Davis and Eells, 1951), but the issue of test fairness and test bias have received more attention from both measurement researchers and the general public since the 1960s. The Senate hearings on psychological measurement and the invasion of privacy awakened the public to the idea that important conclusions affecting one's future could be made on the basis of one's response to certain test questions. The issue of how right, fair, or valid a judgment was, in relation to the test performance, became a question which generalized rapidly to include intelligence and scholastic ability tests, and more recently, achievement tests.

What is meant by test bias? There are several operational definitions of test bias, some of which seem to have little relationship to others, yet all are considered alternate ways in which bias can occur. Test bias occurs when a test does not measure the same dimensions across different groups, when a test measures dimensions not sufficiently related to the criterion, or when results are used in selection in an unfair manner. The end result of bias is unfairness and limitation of access for individuals and groups who are deserving of access.

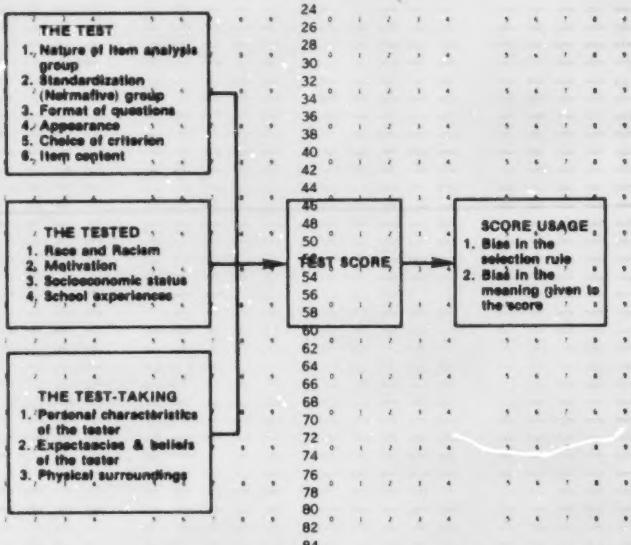
"Test bias occurs when a test does not measure the same dimensions across different groups, when a test measures dimensions not sufficiently related to the criterion, or when results are used in selection in an unfair manner."

These multiple ways in which bias occurs can be considered in two categories; one is related to the test score itself, and the second is related to its usage. These can be further divided as

follows: First, the test score itself can be affected by psychometric factors built into the test itself, by characteristics of those groups being tested, and by aspects of the testing setting. Secondly, test use can occur in selection for schooling or employment, as well as the meaning and interpretation given to a score. These sources of bias are shown in Figure 1, and will be discussed in the following sections.

Figure 1

SOURCES OF TEST BIAS



A. PSYCHOMETRIC FACTORS IN TEST BIAS

There are several steps and decision points in the process of constructing a test, and any juncture along the way is a possible source of test bias. A test plan is developed which outlines specifications for the test questions, or for sets of questions. A question on a test is called a test item. The specifications may include the type of content the item should contain, the type of process to be involved in solving the item, the desired difficulty of the item, the format, and any other item characteristics considered important.

After a set of items has been written, discussed, and edited, they should be tried out on subjects similar to those for whom the test is intended. This is not the norming, or standardization process, but is an important step because it helps to determine which items will actually appear on the test. In this tryout process, the hunches of the writers and editors are supported or refuted by how well the available subjects actually perform.

Often the tryout items are administered in conjunction with an actual previously developed test which serves as a yardstick for judging the quality of the new items. If subjects who do well on the "old" test do well on the tryout item, and subjects who do poorly on the "old" test do poorly on a tryout item, we say that the item discriminated well, or has a high discrimination index. (This is discrimination in a descriptive sense, not in the sense of racial or sexual discrimination). Such an item is said to "work", and is likely to become a part of the test being developed. On the other hand, if subjects who do poorly on the "old" test do well on the tryout item, we say that the item discriminates negatively, and the item is likely to be rejected, revised, or the "correct" answer changed. Often there is something wrong with such items in terms of meaning, style, or some other factor.

Another category of tryout items may occur. These are items that do not discriminate in one direction or the other. That is, subjects who do well on the "old" test perform on a certain level on the tryout item, and subjects who do poorly on the "old" test perform at about that same level. By many test developers, such items are omitted from the test being developed, since they do not appear to add to the information to be gained from the test. These may be called "flat" items, due to the shape of the item characteristic curve, a graph of item performance relative to total score performance.

This group of "flat" items can be quite an amalgam. It may include some poor or misunderstood items, but it may also include items with good discrimination power for subgroups for whom their content is appropriate, and low discrimination for others. Items with greater validity for minority group members can easily be eliminated in this tryout process, particularly if the proportion of minorities in the tryout group is small. A test that has been well-constructed in some ways, such as careful attention to item format and content, may be seriously biased if the items selected are chosen after tryouts with groups containing few minorities. Some developers do not depend as heavily on tryout data in item selection, but place more emphasis on careful, logical selection by experienced editors and writers. Yet these item writers may be very similar in background and life experiences, as well as in training and academic experiences and thus become an additional source of bias themselves.

1. Item Content as a Source of Bias

The specific subject matter content of test items may make them inappropriate for minority groups, because of unfamiliarity with prerequisite material, or because they convey different meanings for some groups. Even though the necessary background in formulation is contained in the test item, previous contact with that information may be a factor in solving the item.

The relative value of information and reasoning skills in solving a test item problem is an important one from the standpoint of the purpose of the test. Consider a test question from the verbal section of a major professional aptitude examination, where the item content is related to European history. To answer the item correctly, one had to recognize that Great Britain's national defense strategy had historically emphasized naval strength, in part due to its being an island nation. The test item included a map, and made reference to countries on the European continent.

This type of question could be answered correctly by logical reasoning and inspection of the map, even without previous knowledge of European history and geography. However, it is likely to be a much easier item for the student who has the substantive background in history. If one can imagine two hypothetical individuals whose developed abilities are similar, but one's academic experiences make a test item easier for that person, then that person's score is likely to be

higher than that of the other. If the purpose of the test is determining level of achievement, then this is a valid item. If the purpose instead is to measure aptitude for a graduate or professional program, it has less validity.

Now aptitude test items cannot, and perhaps should not, be content free, particularly at advanced levels. By sampling a variety of content areas, it is assumed that the advantages of the type in the previous example are roughly equalized. Somebody gets a break because they know something about European history, someone else gets a break on another question due to a background in political science, and an elective in art history gives a boost to another. But suppose areas of academic study and interest are not sampled? A study by Medley and Quirk (1974) gives an insight into what can happen. Beginning with a sample of examination questions from the National Teachers Examination (NTE), they developed a parallel set of items which differed only in content from the original set. The new set of items contained content related to Black history, institutions, and individuals. The reading level, and type of analysis, interpretation, and reasoning required by the two sets of items were the same. Both sets of items were tried out on samples of Black and white college students. The Black students performed better than the white students on the Black-oriented content items, and the white students performed better on the "standard" content items.

The issue of content bias may seem relevant in the "soft" sciences and humanities, but how about the natural and physical sciences? Shouldn't they be bias free? There is an intuitive feel of rightness about the conjecture that if two individuals have taken the same science courses, and performed well, they should know the same thing. In fact, the same science courses at different institutions do vary, and the validity problem may be a greater one than in the case of social science material. After all, a strong analytical mind with limited background in European history may deal effectively with verbal items based on such content, because it can apply careful reasoning to the material given in the item, but it can be totally ineffective in handling organic chemistry items without the necessary academic background.

But then, isn't this the academic background necessary for the study of medicine? Up to a certain point, yes. But remember that we're talking about a test designed to discriminate among a select and bright group of people. Such

items may deal with scientific material, but in a more convoluted fashion.

For example, several years ago, one of our major state universities initiated a special scholarship program for students who scored above a 650 average on the College Entrance Examination Board Scholastic Aptitude Test (SAT). A new test was devised to discriminate within this select group. It was found that two types of items did the best job of "spreading out" this group. The first type of item was characterized by several sequential steps, and required sustained effort in a chain of processes of increasing difficulty which were similar to Bloom's taxonomic levels (Bloom et al, 1956); this chain involved using knowledge, making an interpretation, application, analysis and synthesis. These items also required a flexibility of thought, and ability to "shift gears" as one proceeded through a task.

The second type of item simply employed obscure and archaic vocabulary. Many students got this second type of item wrong, but whether this performance is really related to the criterion of scholastic aptitude is questionable. The item discrimination power was equal to the first type of item, but clearly with much weaker validity in relation to a conceptual criterion. Using item-test correlation, a traditional measure of item validity, these vocabulary items appeared to be as valid as the first type of items (S. Johnson, 1963).

A word should be said about the format of the test item, which can be considered one aspect of content. Standardized test items are generally multiple-choice, paper and pencil tasks. There are ingenious variations in style, but this basic format, that is easily processed by computer, remains. That commercial testing might become locked in such a format was a concern of the late inventor of high speed test scoring and processing, E.F. Lindquist, and it appears that his fears were well-founded (ACT, 1978).

2. Subject Characteristics as Factors in Test Bias

a. Race, Racism and Test Bias

Classic works on the effects of race on test performance was done by Katz et al. (1972) and Robinson, Epps, and Waly (1964). Through a set of experimental conditions involving Black and white examiners with Black college students, the findings of these studies strongly

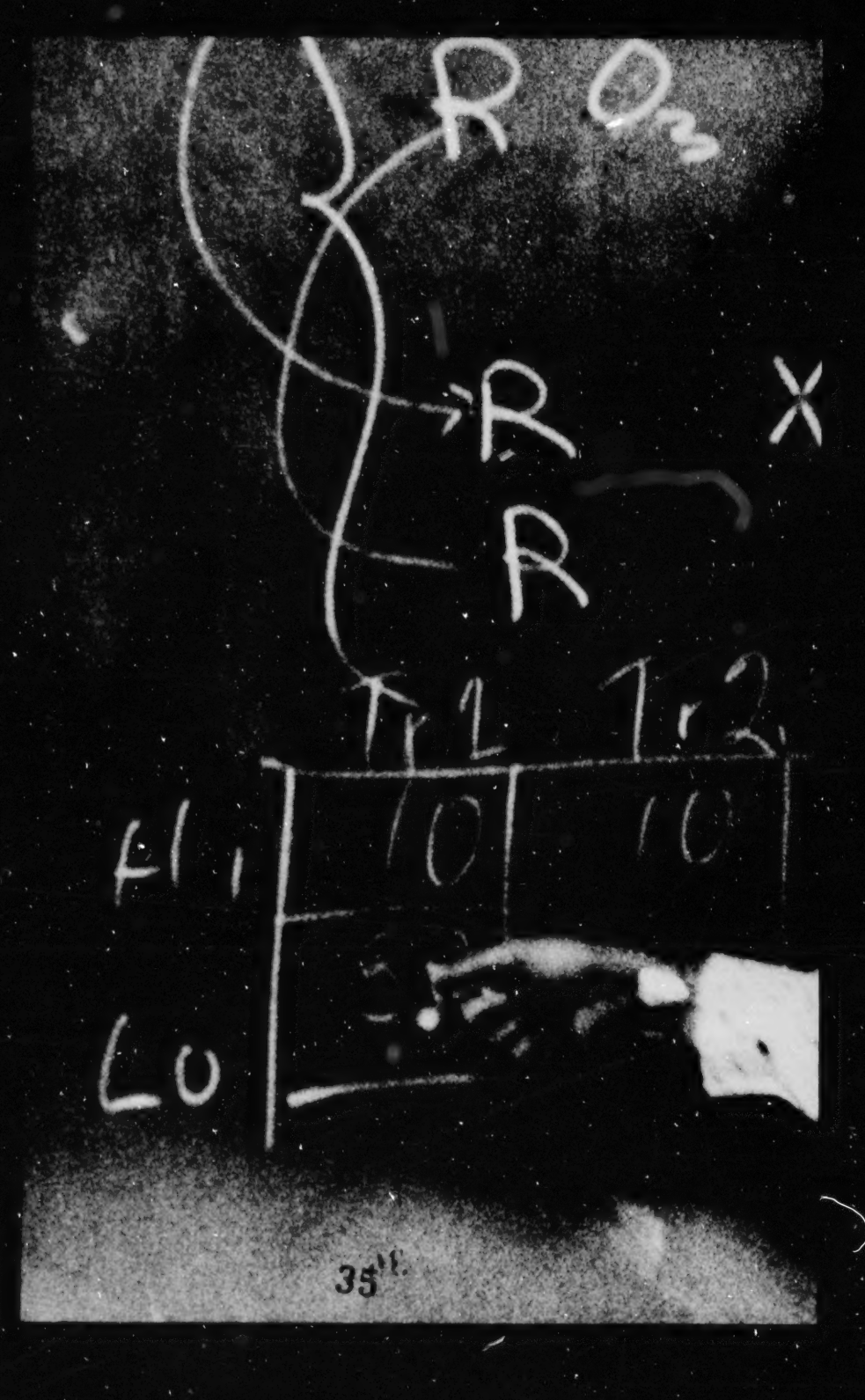
suggested that race of the experimenter, or some characteristic highly confounded with race, functioned to depress digit-symbol performance test scores of Black subjects, and to arouse hostility and anxiety in them. Subsequent studies by these and other investigators have varied race of experimenter, race of subject, socioeconomic factors, and other variables (Katz et al., 1972; Epps et al., 1972). An old, but recycled position has been taken by Jensen, Herrnstein, and others that test performance of the races differ due to genetic differences (Jensen, 1969; Herrnstein, 1971). This position is adequately considered elsewhere (Williams, 1976).

In fact, what is called race in these studies may be more accurately and meaningfully defined as racism. In those studies that show performance decrement in Blacks when the examiner is white, it is not the race of the examiner as such, but the meaning that the race of the examiner has for the subjects due to their previous experiences, or the meaning that the race of the subjects has for the examiner. This seems obvious, and has been pointed out many times before. It seems important to stress, however, because racial diversity is a fact that will not change; the existence of institutional racism is also a fact, but may be amenable to change. Focusing on race as a "cause" of performance differences, rather than the fact that we perceive race in this country as we do, is bound to be counterproductive.

b. Motivation as a Factor in Test Bias

Aptitude and achievement tests are commonly described in measurement texts as "Tests of Maximum Performance." That is, the quality of their estimates depends on the subject's performance being the best that he/she can do at the time of the test. If subjects are not working at maximum performance levels, the test is not obtaining accurate information. It is information that is both unreliable and invalid. To make the inference of maximum performance, the motivation of the subject should be considered; that is, what is the orientation of the subject toward achievement on the task? Of course, anyone who takes an admissions examination for graduate or professional school ought to be seriously interested in the task at hand. Yet, consistent, intensive involvement with the test task for the three to six hour period required may be affected by intrinsic and extrinsic motivational factors.

This complex area of motivation has been extensively addressed in the literature from various theoretical frameworks. A recent postulation by Banks et al. (1978) seems


R
O_3
R
X
R
Tr1
Tr2
Fl
10
10
LO
35"

to offer the promise of integrating previously conflicting findings. Banks presents empirical findings in support of the conclusion that interest in a task strongly affects intrinsic achievement orientation toward the task for both white and Black students. This common-sense conclusion, supported by a series of experimental studies, takes on special significance when viewed in the context of traditional research studies which have found Blacks lower in intrinsic motivation, but have employed tasks which were of less interest to Blacks. The role of interest in reading comprehension test performance has been previously reported, and is consistent with Bank's formulation (Estes and Vaughan, 1973).

Interpreted in light of conclusions by Banks et al., earlier studies of test tasks take on new meaning. The Medley and Quirk (1974) study discussed earlier in this paper should be viewed in the context of the Banks' findings. Recall that in that study, two sets of test items were constructed on the National Teachers Examination, one with content of interest to Blacks, and the other with content of interest to whites. Each group did best on the items with content geared to them. This important factor of interest in test content has been given limited consideration by test publishers, and offers real potential if the findings are incorporated into test construction practice.

Studies in which differing performances are found for different types of tasks should be examined in light of the Banks' findings. Littner and Ducette (1974) found interaction effects between type of motivation and previous failure experiences for an ambiguous coding task but not for a school-related task. Examination of interest factors on these tasks might offer information here. Also, Flaugher (1971), in a study of *Project Access*, examined scores of Caucasian, Black, and Chicano students on traditional and non-traditional measures. He found extensive racial differences on traditional measures, but much smaller differences on non-traditional measures. Differences of this type are sometimes attributed to low reliability of the "new" instruments, but the measure showing the least difference had a reliability of .93. Validity of such non-traditional item types can be explored, and these items included in typical aptitude tests. Also, if items of these types seem to be less typically influenced by socioeconomic differences, the relationship between these subtests and more "traditional" tests may be a potential indicator of the educational disadvantage. Such a use would have to be highly cautious, recognizing the "Sputnik leaps" that a more economically able and responsive school system is able and willing to make when its students are perceived to be placed at a disadvantage.

The role of school differences and how these differences interact with motivation is a complex one that is related to test performance, but will not be discussed here. When racial groups score more alike on a particular type of item, it is likely that this is an item that has not received much attention in the schools. The effect of curricular emphasis on test scores has been reported as a major reason for the high correlation of verbal and mathematical subtests on aptitude and achievement tests (Thorndike, 1971). It was proposed that the apparent low correlations of the Iowa Tests of Basic Skills (ITBS) map reading subtest scores with other subtest scores was because these items measured skills which at that time were not being systematically taught in the public schools. This suggested that as instruction caught up with measurement,

---

"There is a substantial body of research to support the thesis that cues from the testing situation, the tester, and the test itself may serve to elicit a response, e.g., anxiety, from some subjects that interferes with their attendance to and performance on the test task."

---

this subtest would correlate more highly with verbal and mathematical subtests. The fact that traditional subject matter and ways of measuring do often initially reduce group differences strengthens further the thesis that these differences are a function of educational opportunity.

Some measurement specialists discount bias related to motivation. Ebel (1975) has indicated, "If a pupil does not choose to have his educational achievements measured, no test or no tester can measure them. His score will be biased, but it is not the kind of bias that can justly be blamed on the test or the tester. Only himself and his cultural background are to blame" (p. 7).

Such a statement ignores the fact that the nature and impact of the test and tester directly affect the pupil's "choice." There is a substantial body of research to support the thesis that cues from the testing situation, the tester, and the test itself may serve to elicit a response, e.g., anxiety, from some subjects that interferes with their attendance to and performance on the test task. Another body of work supports the conclusion that social-motivational variables, such as self-concept, locus of control, achievement motivation, and expectancy interact with cues of the test and testing situation to raise or lower performance. There is evidence that these variables are relatively enduring, though modifiable, charac-

teristics of individuals formed as a result of feedback from the environment. If you will allow a considerable condensation of the process, one could say that if individuals get many messages saying that striving for achievement involves rejection and pain rather than enhancement, then they will behave accordingly in a situation that arouses these motives. Many aspects of society give these messages more often to Black children that come from powerless low socioeconomic status (SES) backgrounds than to children from middle-class backgrounds (S. Johnson, 1974). Conditions that alter these social-motivational variables have been shown to have positive effects on achievement (Haynes, 1978).

The range of test scores achieved by Blacks and other racial minorities is equal to that of whites on the graduate and professional examinations. Thus, there are numbers of Blacks in every test score category, from highest to lowest. Yet, Blacks are aware that test scores and scholastic achievement have not resulted in success for large numbers of their role models. The bare facts of discrimination against Blacks in schools and employment are a common part of the background that a Black child "grows up knowing," and they are a factor in motivation in any testing situation. These facts do not really need documentation, but it has been noted that in 1955 a number of medical schools still would not admit Afro-American applicants at all (Raup and Williams, 1964; Cuca et al., 1976). Instead, a status dating back for 50 years as the major employer of college educated Black men has been held by the U.S. Postal Service and its predecessor, the Post Office, rather than the range of private and public endeavors which should have made use of this broad talent. It is only during the past fifteen years that any range of options for employment have been available for this group. Even in 1963, Robert Weaver noted that Black males lost substantial ground in income relative to white males of the same educational level as they rose up the educational ladder. This picture is changing for those Blacks and minorities with advanced degrees and high research productivity (Freeman, 1976). This may be truer for the outstanding, rather than the average college graduate (Walster, Cleary, and Clifford, 1970).

Thus, even if one wants to do very well, the degree to which performance is maximized is bound to be related to how much good one thinks it will really do. Many Blacks have had strong reason for questioning whether it is really to their benefit to perform well on a test.

c. Socioeconomic Status

One of the most enduring factors affecting test scores is socioeconomic status. Numerous studies have shown size of test score to be related to parental income level. Data for the MCAT, 1975-76 is in Table 3. This table shows a rank order correlation of .95 between test scores and income for each of the four MCAT scores.

Table 3

MEDICAL COLLEGE ADMISSION TEST SCORES, UNDERGRADUATE GRADE-POINT AVERAGES, AND ACCEPTANCE RATES OF APPLICANTS BY PARENTAL INCOME, 1975-76 FIRST-YEAR CLASS *

Parental Income	Applicant Pool		Mean MCAT Scores +				Grade-Point Average	Percent Accepted
	Total No.	Percent	VA	QA	GI	Sci		
Less than $5,000	1,652	5.3	501	540	489	514	3.09	33.8
$5,000-9,999	3,156	10.2	523	562	511	543	3.20	36.3
$10,000-11,999	2,981	9.6	535	576	518	559	3.25	38.5
$12,000-14,999	4,059	13.1	539	582	525	566	3.28	38.4
$15,000-19,999	5,024	16.2	546	592	530	578	3.31	41.5
$20,000-24,999	4,640	15.0	553	597	535	585	3.31	43.5
$25,000-49,999	6,063	19.6	560	601	540	588	3.29	46.0
$50,000 or more	2,997	9.7	557	596	538	582	3.25	47.9
No response	412	1.3	562	578	541	568	3.22	36.8
Total ‡	30,984	100.0	544	586	528	571	3.27	41.6

*Includes only those applicants to the 1975-76 first-year class who took the MCAT in 1974—approximately 73 percent of the entire pool of 42,303. Incomes reported are for 1973.

+ VA = Verbal Ability. QA = Quantitative Ability. GI = General Information, and Sci = Science.

‡ The median parental income was $18,400.

Note. From "Study of U.S. Medical School Applicants, 1975-76" by T.L. Gordon and D.G. Johnson, *Journal of Medical Education*, September 1977, 5, 707-730. Reprinted by permission.

Apparently conflicting results are reported in a study by Waldman (1977) of applicants in 1976-77. Waldman reported minimal differences in test scores and Grade Point Averages (G.P.A.) on the basis of parental income level. However, the use of only two income categories, and failure to recognize the level and skewness of the parental income curve for Blacks may well account for those results.

3. Setting and Administration Factors in Test Bias

a. The Tester

The tester is a possible source of bias. A large body of literature attests to the effects of warmth, familiarity, trust, and other personal characteristics of the tester on test performance, particularly with minority group children. A study by Candelori in New Haven, Connecticut in 1972 employed personal warmth of the tester as an independent variable. Groups of Black and white children were randomly assigned to "warm" and "cold" testers. The "warm" testers behaved in a friendly, supportive fashion, while the "cold" testers were distant, aloof, and abrupt. Black children performed significantly better under the "warm" testing situation. Of particular interest is the fact that an interaction between race and social class occurred. That is, within the Black group, tester warmth had a marked effect of low SES Blacks. Within the white group, warmth had some effect, but it was not significantly greater for low SES white children than

---

"The tester is a possible source of bias. A large body of literature attests to the effects of warmth, familiarity, trust, and other personal characteristics of the tester on test performance, particularly with minority group children."

---

for high SES white children. This means that the warmth of the tester (and possibly other related interpersonal characteristics) had its greatest effect on the performance of low SES Black youngsters. Thus, in areas where there are large proportions of such children, performance may be significantly affected by such tester characteristics.

The needs and expectancies of the tester are important potential sources of test bias, but their effects may not be as easily identified as in the case of observable tester behaviors, such as warmth. The literature supports the role of examiner needs and expectancies as factors which affect the motivation of the person tested. Shafer (1952) has pointed out in detail how these personal factors may operate in unobservable ways to influence testing outcome, either by causing the subject to respond differently to the tester, or by causing the tester to interpret the responses given by the person being tested in a way that results in a different or lower test score. This motivation, then, can significantly affect test scores, since in order for the test score to give any valid estimate of maximum performance, the person being tested must be motivated to perform.

The effects of tester needs and expectancies may be transmitted in ways other than by apparent warmth or coldness of the tester. The tester is responding to both internal cues which include his or her own needs and to external factors, such as the source of the testing referral. The tester's feeling about the person or agency making the referral for testing may affect her/his behavior in the testing setting. Hersh (1971) noted that the clinical testing situation forms a triad composed of the referral agent, the tester, and the testee. In his research, the Stanford-Binet Intelligence Scale was administered to 56 Headstart program children. Each examiner tested both a child supposedly referred for "positive" reasons, and one referred for "negative" reasons. "Positive" reasons included gathering information for academic planning, and identifying areas of competency. "Negative" reasons included disruptive behavior in the classroom, and conflicts outside of school. Those children with positive referrals obtained higher I.Q. scores, were rated more favorably on factors affecting test performance, and received more favorable recommendations than for children whose referral statements were negative. Yet, no significant differences in tester warmth were observed.

b. Physical Surroundings and Bias

It is generally accepted that testing conditions; i.e. noise, light, space, etc. must be suitable, or test results will be invalid. In fact, the physical surroundings can be much more important, because they provide cues to the person being tested. The kind of message given by a down-trodden setting that shows little hope of improvement may give that precise meaning to the test taker. Expectation can be a powerful motivator, and some settings give a message loud and clear.

B. Bias Resulting From Improper Test Use

Improper use of test results is another source of bias. This can occur by the attribution of unwarranted meaning to the results, as in using a test score as an indicator of "innate potential." It can also occur whenever unwarranted or invalid usage is made of a test. It would seem that increased information about tests would result in better test use. However, Buros (1965) observed:

> Unfortunately, the rank and file of test users do not appear to be particularly alarmed that so many tests are either severely criticized or described as having no known validity. Although most test users would

> probably agree that many tests are either worthless or misused, they continue to have the utmost faith in their own particular choice and use of tests regardless of the absence of supporting research or even of the presence of negating research. When I initiated this test reviewing service in 1938, I was confident that frankly critical reviews by competent specialists representing a wide variety of viewpoints would make it unprofitable to publish tests of unknown or questionable validity. Now, 27 years later and five *Mental Measurements Yearbooks* later, I realize that I was too optimistic. Although many test users undoubtedly are selecting and using tests with greater discrimination because of the MMY's, the publication and use of inadequately validated tests seem to be keeping pace with the population explosion. (p. xxiii)

The responsibility for correct test use also lies with test developers. Green et al. (1975), notes:

> While it is basically true that the test user should validate a test on each occasion of its use, the test maker has the primary responsibility for making this clear to the user. Just as the television viewer will only watch the current programming fare, the test user, who may not be sophisticated in test theory or the use of tests, cannot go beyond the information he is given by the test maker. When test makers do not firmly establish the quality of their instruments, they may reveal that their primary concern is with profit-making. (p. 4)

A frequent misuse of tests in educational settings is in using tests to classify and assign children on a long-range basis. This usage has resulted in low achievement in depriving minority and poor children of educational experiences which might have prepared them to take on even more challenging experiences, and has limited their educational horizons (Green et al., 1975; Kozol, 1967). A further consequence of this misuse is the interpretation of the resulting low achievement gains for minority children as a indictment of desegregated schools. This is a conclusion made by Jencks (1973) although he did note the positive effects of desegregation in increasing socioeconomic status.

A new area of test use with potential for misuse is the requiring of exit examinations for elementary and high

school. This is a new area, and usage appears to be growing faster than careful research would warrant. Possible consequences are the limitation of what is learned, rather than the expansion and growth of the full potential of all individuals.

1. Bias in the Selection Rule

Bias in the selection rule itself which is used for selection to academic or employment programs can be considered the most important area of bias, since without gaining access to education and employment, the whole process of the removal of the long term effects of discrimination is thwarted. One major issue in the recent literature on test bias is the concept of group versus individual parity. The group parity concept involves the evaluation of tests and test items in terms of their fairness to different groups, while the individual parity focus is on fairness to individuals. Using various theoretical approaches, a series of models have been developed by measurement specialists to identify and correct bias in selection. Both individual and group parity approaches begin with individual fairness as a goal. The group-related parity concept grows out of the fact that systematic unfairness, such as the consequences of racism, results in group-related unfairness, and therefore requires group-related remedies. The group parity models regard group parity as a natural outcome for a system based on individual merit, when a group of individuals has been dealt with by systematic unfairness. This evolvement of a group-related system as a function of systematic unfairness is presented by Sawyer, Cole, and Cole (1973). An individual parity model based on a conditional probability approach is represented in the work of Petersen and Novick (1976).

In an earlier article, Cole (1973) pointed out the need to employ models which have been used to identify bias as models to insure fairness. That study analyzed the major bias models in use at that time, and demonstrated what effect their use would have on selection of minorities and majority groups. Situations of different types of differential validity that had been documented in the professional literature were examined.

Cole then applied each of the six models designed for fair selection to each situation of differential validity. Five models had been previously developed by other researchers; the sixth, the conditional probability model, was developed by Cole to meet some of the limitations of the other models.

The conditional probability model was designed to provide fairness in selection to a group of applicants who, if selected, would be able to succeed. The basic principle of the model is that there should be the same probability of selection, regardless of group membership, for both minority and majority groups whose members can achieve a satisfactory criterion score. That is, the model is designed to provide fairness in selection to all who can function effectively as physicians, lawyers, or whatever the criterion of interest is.

The Cole article elicited tremendous response within measurement circles, with the result that a new model for fairness was proposed (Petersen and Novick, 1976), and additional ways of approaching fairness models were suggested (Cronbach, 1973).

Intrinsic to the newer models is a difference in the weighting of the test score or other predictor. This difference is determined by group status, or by some arbitrary function assigned to individuals. In other words, these models recognize that fairness may involve accepting a different meaning for the same test score when it is received by individuals belonging to different groups.

Petersen and Novick (1976) compared and evaluated the major bias models, including Cleary (1968), Thorndike (1971), Cole (1973), Linn (1973), Einhorn and Bass (1971), and Gross and Su (1975). They conclude that models such as

---

"If decision-makers feel that it is useful and important to have fair numbers of Black and female physicians, then a utility function would be chosen for use in selection. This function would give weight to the minority or sex status of applicants. A different utility function would be constructed if other priorities were present."

Thorndike, Cole, and Linn derive from a concept of group parity and have "highly objectionable practical implications." They then reject as unworkable the concept of culture-fair selection. Instead, they propose an expected utility model, based on a utility function which would be determined by "decision-makers in conjunction with all interested parties (who) openly and publicly decide upon appropriate utility functions." The utility functions would be made according to existing needs and pressures, and could be affected by desire for fairness to underrepresented minorities, or other circumstances. If decision-makers feel that it is useful and important

to have fair numbers of Black and female physicians, then a utility function would be chosen for use in selection. This function would give weight to the minority or sex status of applicants. A different utility function would be constructed if other priorities were present.

In the concluding article of the Spring 1976 issue of the *Journal of Educational Measurement,* Novick and Petersen further reject the concept of group parity, while accepting the Sawyer, Cole, and Cole (1976) thesis that, if there is differential predictability for two groups and if prediction is much better for the advantaged group than for the disadvantaged group, then, the use of that instrument will further disadvantage members of that group. The Novick and Petersen solution is to depend on the evaluator's sensitivity in setting utilities, and more recently to develop an "index of disadvantagement" (Novick, 1978) which could be determined for any individual and used to determine the utility function to be employed for that individual.

The Sawyer, Cole, and Cole (1976) discussion does not begin with the notion of group parity, as Novick and

---

"Tests, as predictors of criterion measures, do have value in selection. Without any test scores, selection might be a very haphazard, subjective process, and unfair to many. The problem is in the implicit substitution of the test score for the real thing that we want to measure."

Petersen imply. Rather, it concludes that a system based on individual merit becomes a group related system, when identifiable categories of people produce different prediction systems. When uniform unfairness has been the rule, it is not surprising that a fairness solution has a strong group related component.

2. Bias in the Meaning Given to a Test Score

A number of factors affecting the performance of Blacks and minorities on tests have been discussed in the preceding sections. The question of what meaning is to be ascribed to the score must then be considered. The current achievement aspects of a test score may be useful, but there is a risk of over-interpretation. This might be termed "bias by belief," and is a result of blind adherence to the sanctity of the score by applicants and the general public, but less so by the academic and commercial users of tests.

Flaugher (1978) points out that a low score seen as a measure of aptitude can be used for withdrawal of educational resources, while the same result seen as an index of achievement should work toward increasing the resources to raise achievement. Clearly, this results from the interpretation and meaning given to a score.

C. Validity and Test Bias

Tests, as predictors of criterion measures, do have value in selection. Without any test scores, selection might be a very haphazard, subjective process, and unfair to many. The problem is in the implicit substitution of the test score for the real thing that we want to measure. This results in a sort of "validity by definition" position that assumes that the best scores on the test will always be made by those who will be the best doctors, and the worst test scores will always be made by those who will be the worst doctors. The constructors of tests make no such wild claims, although they could do more to make their instruments less misused.

In fact, the term "validity" has a simple, straightforward meaning. By validity, we mean the usefulness, or utility, of a test for a particular purpose. Predictive validity indicates the degree to which the test score is related to success on some future activity or task. We call this future activity or task a criterion. A test has high validity if good scores are earned by people who do well on the future criterion task, and poor scores are received by people who do poorly on it. When we use tests to select people for a school or training program, we are interested in predictive validity. We need to know how well the test helps us in deciding whether an applicant will do well or not. For instance, we are interested in predictive validity when we give kindergartners a reading readiness test. Our criterion task is achievement in first grade reading. Predictive validity is also of concern when a clerical test is administered to job applicants. How good they will be as clerks after six months of service is a criterion of interest.

When we determine the validity or "goodness" of a test for a particular purpose, it is also important that the validity be consistent for the people on whom we intend to use it for selection. This is more likely to be the case when a group of people have similar backgrounds. That is, when they are similar on variables other than the criterion of interest.

When a Medical Admissions test has good predictive validity, there will be some people with good scores

who will make good grades, and some who will make poor grades. We might call this first group "hits" and the second, "misses". Also, among those with low test scores, there will be some, if they got into medical school, who would make poor grades and some who would make good ones. In terms of the validity of the test, these potentially poor grade-earners are "hits," and these potentially high-grade earners are "misses," since these individuals with low scores would probably not be admitted to medical schools. A certain amount of error in selection is inevitable, with or without tests, even when the pool of subjects is similar.

In contrast, when the members of a group from which we intend to select come from widely different backgrounds, that is, they differ in many respects other than the criterion of interest, it raises additional problems for selection on any basis, either by the use of a test or other methods, because we are likely to encounter the phenomenon known as *differential validity*. In the *Bakke* case, the academic and cultural backgrounds of applicants varied widely. From the information we have on those minority students who have succeeded, it appears that there is also differential validity.

By differential validity, we mean that the relationship between test score and criterion is different for different groups. In this case, the relationship between traditional predictors and success in medical school is not the same for minority students as it is for majority students. Thus, the meaning of the test score itself is different for the two groups.

The test is identifying different proportions of "hits" and "misses" in the two groups, and the classification of high scorers into success and failure groups differs for the two groups. Also, the classification of low scorers into success and failure groups differs for the two groups. When differential validity exists, it is unfair to apply the same selection process to all applicants, because it would discriminate against minority applicants.

Since we tend to equate "the test" with the quality, characteristic, or skill that we are trying to measure, the above reasoning seems at first to "rub against the grain", from a logical frame of reference. After all, isn't the one with the best test score the one who will be the best doctor, or the one whose knowledge is greatest? Even under optimum circumstances, this is a difficult question, since we must agree on what "best" means, and what knowledge we are talking about. With an uneven relationship between predictor and conceptual criter-

ion, or with a situation of differential validity, the answer to this question can well be no.

"After all, isn't the one with the best test score the one who will be the best doctor, or the one whose knowledge is greatest? Even under optimum circumstances, this is a difficult question, since we must agree on what "best" means, and what knowledge we are talking about."

When we are choosing from people with similar backgrounds, it is assumed that the relationship between a well-developed selection instrument and the criterion (in this case, performance in medical school) is the same for different randomly selected samples from the pool. Thus, we can with some confidence make the same estimate for degree of success in medical school for two individuals whose test scores are the same. In the real heterogeneous world, however, the meaning of the test score for different individuals can be quite different. When someone applies for a job or for professional school, we cannot know how well they will perform on the criterion.

If an arbitrary level on the Medical College Admission Test indicated poor promise as a physician, we would have cause for concern. The average verbal ability score of students admitted to medical school in 1955-56 is virtually the same as the average score of those rejected in 1975-76. The average quantitative ability and science scores of students admitted in 1955-56 is 17 to 34 points lower than the scores of students rejected in 1975-76. The lower scores of 1955-56 are not due to minority students, since they constituted less than 3 percent of all medical school enrollment then, and they were 6.15 percent of the admitted students in 1975-76. These changes in scores of admitted and rejected students are shown in Table 4.

D. Reliability and Test Bias

While the thrust of this paper is on validation problems, the reliability of aptitude tests must also be considered. Reliability is the consistency or stability of measurement. To the extent that what we are measuring behaves consistently on our measure, the measure is reliable. To the extent that what we are measuring varies randomly, the measure is unreliable. Thus, reliability can be defined as a function of the proportion of "true" and "error" spread, or variance in scores. We can measure reliably, and at the same

Table 4

MEAN MCAT SCORES OF ACCEPTED AND REJECTED APPLICANTS

First-Year Class	Verbal Ability	Quantitative Ability	General Information	Science
		ACCEPTED APPLICANTS		
1955-56	524	528	527	522
1960-61	527	533	527	533
1965-66	541	583	565	549
1970-71	559	606	560	558
1975-76	575	620	550	615
		REJECTED APPLICANTS		
1955-56	466	459	476	454
1960-61	464	453	473	449
1965-66	473	502	511	466
1970-71	506	539	518	499
1975-76	522	562	513	539
		DIFFERENCES BETWEEN ACCEPTED AND REJECTED		
1955-56	58	69	51	68
1960-61	63	80	54	84
1965-66	68	81	54	83
1970-71	53	67	42	59
1975-76	53	58	37	76
Average	59	71	48	74

Note. From "The Medical School Admissions Process. A Review of the Literature, 1955-1976." *Final Report of Contract #231-76-0011 to U.S. Department of Health, Education and Welfare, Public Health Service, Health Resources Administration* by Cuca, J.M. et al., Rockville, Maryland: National Technical Information Services, October, 1976. Reprinted by permission.

time do invalid measurement. The reverse is not true—if measurement is valid, it must have some consistency or stability. Reliability, through most of the score range, is the upper limit of validity.

The test scores of Blacks, Chicanos, and other minorities tend to have lower reliabilities than those of whites. This may be due to the sources of test bias previously discussed, since many of those arguments also relate to consistency of measurement. It may also be due to restriction of range or the measure being considered, because if the scores

do not extend over a broad range, the ordering scores within the range will be more subject to error.

Reliability can be considered along different parameters. Rather than considering performance of different subjects on test items, we can consider performance of school districts over time, performance of grade levels over school, or other combinations. In each case, error, or random variation, is differently defined. In the growing emphasis on aggregate, or grouped data studies that are done in evaluation of educational programs, careful attention should be given to the way in which true and error variation is defined.

Further attention to the nature of reliability or unreliability in minority test performance may add to our understanding of bias in testing.

III. USING REGRESSION ANALYSIS IN DETERMINING PREDICTIVE VALIDITY

Regression analysis is the procedure used to ascertain the degree to which a person's criterion measure can be estimated if we know a predictor score. If we know nothing at all about an individual, the "best guess" that we can make of how well she will do on the job is to use as an estimate the average criterion score of other people who have applied for the job. This represents maximum uncertainty about the performance of this new person. By applying regression analysis to the criterion and the predictor, we obtain a regression equation. If there is any relationship between the predictor and the criterion, the use of the regression equation will reduce our uncertainty about the performance of the new applicant. It will not give the exact criterion score. It may not be close to the exact criterion score. But if we consider all applicants, the predicted scores will, on the average, be closer* to the actual scores than any arbitrary guess would be. Some scores will be estimated more accurately, and some less accurately. A brief introduction to this regression analysis procedure is necessary at this point. The typical regression picture for the case of homogenous selection pool is shown in Figure 2.

---

* More precisely, the sum of the average squared distances between actual and predicted scores is minimized in regression analysis.

Figure 2

PREDICTED AND ACTUAL PERFORMANCE

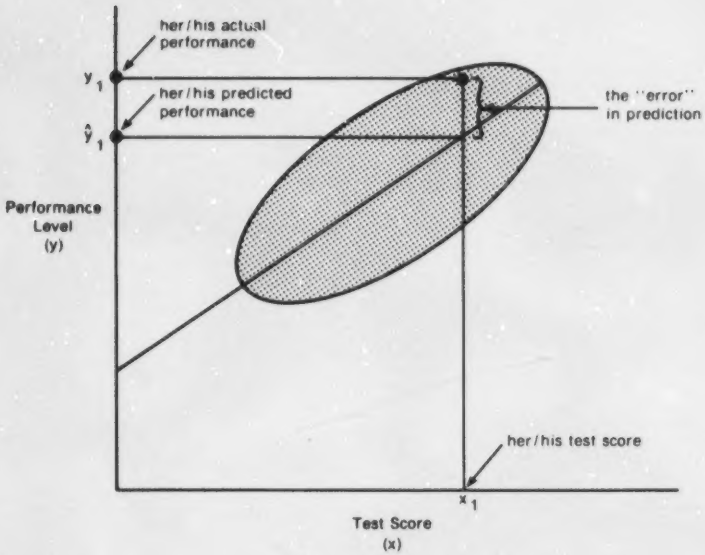


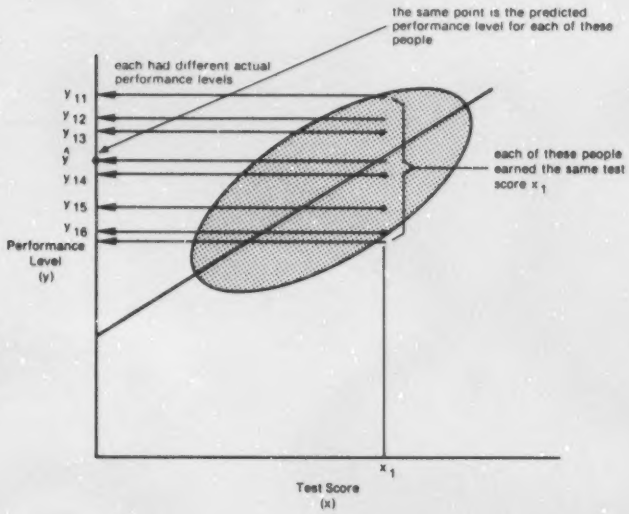
Figure 3 represents the prediction of performance in medical school from test scores. In Figure 3, each dot represents one or more persons. Moving vertically from a dot, we find the test score for that person or persons. Moving horizontally, we find the corresponding performance in medical school or other criterion. In the standard procedure for test validation employing least-squares regression techniques, a regression line is determined in such a way that the average squared vertical distance from each of the dots to the line is minimized. Thus, the distance between the actual and predicted score for each person is minimized.

In this validation procedure, the test being validated is not to be used for selection, that is, the pool of candidates must all be tested, but their test scores are not to be used for selection. They are set aside until the group has completed or failed to complete the program, that is until they have achieved a criterion performance. At that time, the regression analysis procedure is applied to these pairs of test scores and criterion measures, and a regression line is determined. The equation of that line is the regression

equation used to determine the probable degree of success of future candidates similar to those in the original validation group.

Figure 3

ACTUAL PERFORMANCE FOR PERSONS EARNING SAME TEST SCORE



After the validation has been completed, the regression line (or equation) is employed as follows:

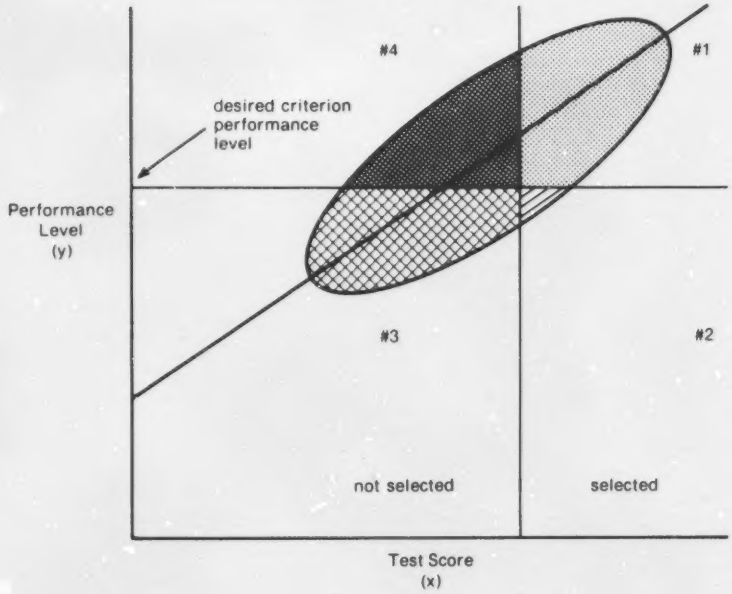
A new applicant takes the test. A vertical perpendicular line is drawn from her/his test score X, on the X-axis, up to the regression line. Then a horizontal line is drawn from that point on the regression line to the vertical (Y) axis. That point on the Y-axis is the predicted Y performance level (Y) for the applicant, based on her/his test score and the test scores for the validation sample. It is the "best guess" that can be made of the applicant's future performance, using least-squares regression.

From the bivariate diagram in Figure 3, it can be seen that all individuals in the validation sample who earned the same X score as did our new applicant did not earn the same Y score. Rather, the Y is comparable to the average Y score of these individuals who received the same X test score.

If the right vertical line determines selection/non-selection, then it is apparent from Figure 4 that of those persons whose test scores were below the selection line, some did perform well on the criterion. These are shown in the shaded area #4 in Figure 4. However, a larger number of low test scorers did not perform well on the criterion (hatched area #3). Of those who did well on the test, some did quite well on the criterion (dotted area, #1). Others did fairly well but no better than many of those who did not pass the test (lined area, #2).

Figure 4

PREDICTED PERFORMANCE LEVELS OF SELECTED AND NON-SELECTED APPLICANTS



In selection, the objective is to maximize the areas 1 and 3, where appropriate decisions are made; those who will do well are accepted, those who will do poorly are rejected. At the same time, an objective is to minimize areas 2 and 4, where unfair decisions are made; people are admitted with high test scores who would perform poorly as physicians, people are not admitted with lower test scores, and these are individuals who in fact, would perform well as physicians.

All of this can be said rather glibly by measurement people, but after all, we have only regression equations, not crystal balls, and we are dealing with probabilities, not certainties. Any applicant with a low test score is certain that her/his score is in quadrant 4, not quadrant 3.

When there is differential validity, the psychometric picture becomes more complicated. Recall that in differential validity the relationship between the test score and criterion performance is different for two different groups. This means that if we were to employ the standard validation procedure, it would result in two different regression lines. That situation is shown in Figures 5, 6, and 7.

Figure 5

PREDICTION DIAGRAM I

Average test performance of two groups is unequal, and predictor has differential validity for the two groups.

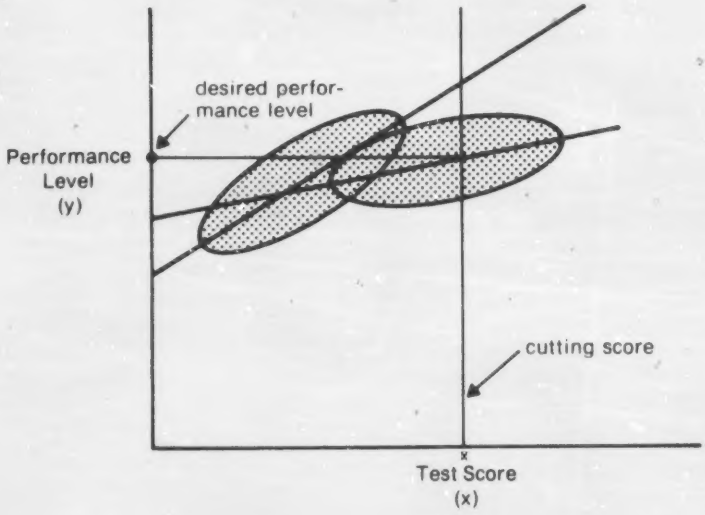


Figure 6

PREDICTION DIAGRAM II

Average performance of two groups is equal, and predictor has differential validity for the two groups

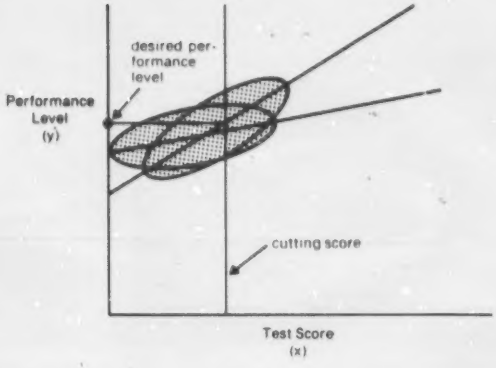
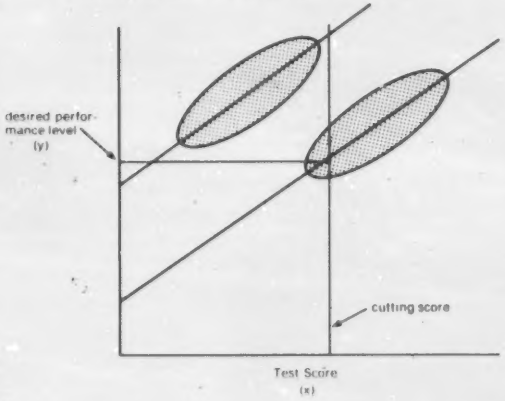


Figure 7

PREDICTION DIAGRAM III

Average test performance of two groups is unequal, and predictor has differential validity for the two groups



In Figures 5, 6, and 7, the use of a single cutting score will unfairly penalize individuals from one of the groups who have an equal chance of success. If Y, a point on the vertical Y-axis, is the desired (not a minimum competency level) level of performance, the minimum test score to be used for selection should vary, depending on the group to which the individual belongs. The use of an inappropriate regression line results in bias, and this is one definition of test bias.

Thus, we see that the error which is inherent in selection, even under the best of circumstances, increases greatly with differential validity, and that a particular test score does not have the same criterion-based meaning for all people taking a test.

This is the traditional least-squares regression prediction picture. In fact, with the emphasis in recent years on test fairness, there are many models for fair selection which recognize the inadequacies of this model. To use them, one must be willing to forego the dream of the absolute or true test score that absolutely predicts the future, and be willing to do a bit of free wheeling in a milieu of uncertainty. Such forays are likely to be met with cries of "reverse discrimination" by believers in the absolute truth of the score.

"One must be willing to forego the dream of the absolute or true test score that absolutely predicts the future, and be willing to do a bit of freewheeling in a milieu of uncertainty. Such forays are likely to be met with cries of "reverse discrimination" by believers in the absolute truth of the score."

IV. THE OVERPREDICTION ISSUE

The concept of overprediction has been explored in several studies (Temp, 1970; Pfeifer and Sedlacek, 1971). In overprediction, the predicted performance is actually higher than the real performance level which the individual achieves. This situation is an artifact, caused by one or more of the following factors: (1) a restriction in range on the test score; (2) restriction of range of the particular criterion measure employed; (3) low reliability of either predictor or criterion; (4) use of a biased criterion; (5) combining of groups of highly unequal size, resulting in slopes and/or intercepts in the prediction equations being different from that of the smaller group; (6) failure to include items with good discrimination power within minority groups. The Pfeifer and Sedlacek (1971) study involved scores of Black and white undergraduates on the Scholastic Aptitude Test (SAT), and their performance in college. They found mild overprediction for Black students when a regression equation based on all entering students was used. They note, however, that the use of this equation reduced the correlation between test and performance from .61 to .46. Further examination of the data reveals an increased standard error of the estimate of grade point average from .52

to .56. Since the standard deviation of the grade point average for Black students was .63, the test provides little benefit in selection.* Noting the existence of overprediction is misleading here. It is not the good quality of the test for use in selection that should be noted. It is the failure to locate characteristics that are really related to achievement for Black students.

Closely related to the "overprediction" issue is the problem of making any estimate from a large group, and comparing it with an estimate made from a small group. For example, suppose you were playing darts, and you had thrown five darts at the target. Suppose then that the "bull's-eye" and the surrounding concentric circles were removed, leaving only the bare board and your 5 darts. How well could you estimate the original location of the bull's-eye? How well could you estimate the average distance from the bull's-eye of all the possible darts you might throw?

Now replace the bull's-eye and target. Repeat your game with 65 darts, again removing the bull's-eye and circles after throwing the darts. Clearly, you can estimate with much greater precision the former location of the bull's-eye. Also,

---

"Because a test had good predictive validity within a certain range of scores, it does not necessarily follow that it has that same validity in higher ranges."

---

your estimate of how far off you are from the center whenever you throw a dart is much better, since you have 65 distances to use to estimate it. The ratio of Black to white acceptances in medical schools in 1975-76 was 1:13, the same as our ratio of darts. Similar problems of estimation occur. Any estimates from the Black sample are based on such small numbers relative to the sample of whites, that the values which they estimate have a larger margin of error.

Because a test has good predictive validity within a certain range of scores, it does not necessarily follow that it has that same validity in higher ranges. That is, because we can

---

* In order for the test to be useful for selection, the standard error of estimate should be much smaller than the standard deviation.

$$s_{y \cdot x} = s_y \sqrt{1 - (r_{xy})^2}$$

say that applicants with scores from 450 to 550 have a much better chance of success in medical school than applicants with scores from 350 to 450, it does not follow that applicants with scores from 550 to 650 have a much better chance for success than applicants with scores from 450 to 550. The test may have performed its most useful function in identifying those applicants who are most likely to fail. However, other predictors may then be more useful in determining who will function best as a family practice physician, surgeon, psychiatrist, or another needed specialty.

V. CONCLUSIONS AND RECOMMENDATIONS

At the beginning of this paper, a set of questions was raised with reference to the *Bakke* case. Let us now re-examine these questions in terms of our discussion.

1. *Is the claim that Bakke's Medical College Admission Test (MCAT) score "entitled" him to admission to medical school valid, from consideration of the test itself?*

   No, the discussions of test bias and predictive validity show that there was sufficient reason to consider a number of characteristics of minority applicants in making a sound admissions decision.

2. *To what extent does psychometric theory support admitting to medical school only those persons having the highest test scores?*

   The regression picture and the MCAT data illustrate that high test scores are related to academic performance in the first two-years, but that within an acceptable score range, other factors may be wisely considered in selection.

3. *To what extent is the use of alternate criteria for admission, (that is, criteria other than test scores and grade point averages) supported by psychometric theory?*

   The model encourages the use of other data, especially considering the criterion problem and the growing number of students with high test scores.

Thus, affirmative action for Blacks, other minorities, and women has in fact a clear psychometric rationale, as well as a moral, sociological, and political one. Fifty years ago, Black college graduates had as major career options service in the U.S. Post Office and the Brotherhood of Sleeping Car Porters. Today, professional employment options are extremely broad, but preparation to exercise these options has been limited by educational opportunity. Part of this limitation in educational opportunity is due to testing, both in the construction and use of tests. In the long run, a system wide approach to the problem must emphasize both: (1) meaningful and productive educational experiences which may be monitored by evaluation instruments, but which must not be stifled by them; and, (2) development and maintenance of fair access to professional education. The problem of increasing the number of minority members in professional schools in the short run may be approached through the following (Breland and Ironson, 1973):

(1) Obtaining more minority applicants.

(2) Tutoring minorities and other disadvantaged applicants in the fundamental skills required for tests, as well as in test-wiseness.

(3) Developing new measures capable of assessing a wider range of competencies.

(4) Creating instructional environments that allow for more diversity in preparation.

Other specific suggestions for development of fair admissions programs are offered in a report of the American Council on Education-American Association of Law Schools Committee on ***Bakke*** (McCormack, 1978). This report outlines several possible law school admissions program criteria and procedures through a case study approach, and appraises their fairness according to the logic of the Supreme Court decision in the ***Bakke*** case. The report recommends several alternate approaches which employ non-test criteria and candidate background information in law school selection, but do not set aside a specific number of spaces for minority applicants. The committee found these approaches fair yet not likely to be in conflict with the ***Bakke*** decision.

There is no argument against the logic that Blacks and whites who seek to master the same professional discipline must each develop the same body of skills and understandings.

What is argued is that the disposition to learn them (aptitude) may be measured differentially with respect to group membership, and this should be considered in selection.

Open affirmative action programs have no need to apologize for using tests as only one part of a selection process. In order to develop a program of fair selection, the weaknesses of employing a purely psychometric basis for selection must be recognized and dealt with squarely. No apologies need be made for including other relevant dimensions in the selection process. The relationship between medical school admissions and adequate health care delivery is not a simple cause and effect system, but these variables are beginning and end points in a complex nomological network. Individual and group related fairness are essential pieces in the network but the end point must remain in focus.

Selection instruments need not be discarded, because they are found wanting. The same argument presented here in assessing tests could be repeated for any arbitrary procedure based on comparing an individual to group performances. The solutions are not psychometric; they are social policy decisions, but they can be made more difficult by psychometric misunderstandings. Psychometricians can give us useful tools, but they must not be misused. This does not absolve test constructors of responsibility. They should be spending as much time and money in assisting selectors for schools and employment in using other predictors as they do in developing tests. In fairness to all, affirmative action has the potential to enrich our society with the contributions to many areas which have too long stood dry from healthy infusion of a diverse racial and sexual populace.

BIBLIOGRAPHY

Arbarbanel, K. "After Intensive Care: Is a Relapse Ahead for Minority Medical Education?" *Foundation News,* Nov./ Dec. 1976, *17*, 24.

American College Testing Program Division of Educational Programs and Services. "Lindquist Translated Ideals into Action." *Activity,* 1978, *16(3)*, 3.

Anastasi, A. *Psychological Testing.* (4th ed.). New York: MacMillan, 1976.

Anastasi, A. "Some Implications of Cultural Factors for Test Construction." *Proceedings of the 1949 Invitational Conference on Testing Problems.* Educational Testing Service, 1950.

Anastasi, A. & Cordova, F.A. "Some Effects of Bilingualism Upon The Intelligence Test Performance of Puerto Rican Children in New York City." *Journal of Educational Psychology,* 1953, *44*, 1-19.

Association of American Medical Colleges. *New Medical College Admission Test Interpretative Manual.* Division of Educational Measurement and Research, Association of American Medical Colleges, 1977.

Astin, A.W. "Criterion-Centered Research." *Educational and Psychological Measurement,* 1964, *24*, 807-822.

Banks, W.C., McQuater, E.V., & Hubbard, J.L. "Toward a Reconceptualization of the Social Cognitive Bases of Achievement Orientations in Blacks." *Review of Educational Research,* 1978, *48*, 381-398.

Barbarin, O. "Race and Social Climate as Determinants of Effective Adaptation in a University Setting." In Cross, W.E. (ed.) *The Third Conference on Empirical Research in Black Psychology.* Washington, D.C.: National Institute of Education, December, 1977.

Barro, A.R. "A Survey and Evaluation of Approaches to Physician Performance Measurement." *Journal of Medical Education,* 1973, *48*, 1051-1093.

Bartlett, C.J., & O'Leary, B.S. "A Differential Prediction Model to Moderate the Effects of Heterogeneous Groups in Personnel Selection and Classification." *Personnel Psychology,* 1969, *22*, 1-17.

Bloom, B., Karathwohl, D., & Masia, B. *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.* New York: David McKay, 1956.

Breland, H.M., & Ironson, G.H. "*De Funis* Reconsidered: A Comparative Analysis of Alternate Admissions Strategies." *Journal of Educational Measurement,* 1976, *13*, 89-99.

Buros, D.K. *The Sixth Mental Measurements Yearbook.* Highland Park: Gryphon Press, 1965.

Cameron, H. "Nonintellectual Correlates of Academic Achievement." *Journal of Negro Education,* 1968, *37*, 252-257.

Candelori, W.P. "The Effects of Non-Cognitive Variables Upon Test-Taking Behavior." *Dissertation Abstracts International,* 1971, *32*, 231.

Carnegie Council on Policy Studies in Higher Education. *The Relevance of Race in Admissions.* Berkeley: Carnegie Council, 1978.

Clark, K., & Clark, M. "The Development of Consciousness of Self and The Emergence of Racial Identification in Negro Pre-school Children." *Journal of Social Psychology,* 1959, *10*, 591.

Cleary, T.A. "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges." *Journal of Educational Measurement,* 1968, *5*, 115-124.

Cole, N.S. "Bias in Selection." *Journal of Educational Measurement,* 1973, *10*, 237-255.

Cooper, J.A.P. "Undergraduate Medical Education." *Advances in American Medicine.* New York: Josiah Macy, Jr. Foundation, 1976.

Cronbach, L.J. "Equity in Selection—Where Psychometrics and Political Philosophy Meet." *Journal of Educational Measurement,* 1973, *13*, 31-42.

Cross, W.E. (ed). *Third Conference on Empirical Research in Black Psychology.* Washington, D.C.: National Institute of Education, December, 1977.

Crown, P.J. "The Effects of the Examiner and Standard vs. Dialect Administration of the Wechsler Preschool and Primary Scale of Intelligence on the Performance of Negro and White Children." *Dissertation Abstracts International,* 1971 (July), 32, 232-233.

Cuca, J.M., Sakakeeny, L.A., & Johnson, D.E. "The Medical School Admissions Process. A Review of the Literature, 1955-76." *Final Report of Contract #231-76-0011 to U.S. Department of Health, Education and Welfare, Public Health Service, Health Resources Administration.* Rockville, Maryland: National Technical Information Services, October, 1976.

Davis, A. *Social Class and Learning.* Cambridge: Harvard University Press, 1952.

Davis, A., & Eells, K. *The Davis-Eells Games.* Chicago: University of Chicago Press, 1951.

Ebel, R.L. *Constructing Unbiased Achievement Tests.* Paper presented at the National Institute of Education Test Bias Conference, Annapolis, Maryland, December 2-5, 1975.

Einhorn, H.J., & Bass, A.R. "Methodological Considerations Relevant to Discrimination in Employment Testing." *Psychological Bulletin,* 1971, 75, 261-269.

Elliott, P.R. "The Admissions Process: How to Select for the Family Practice Student." *Proceedings of the 1975 Annual Workshop on Undergraduate* Education and Family Medicine. Kansas City: The American Academy of Family Physicians, May 1975.

Epps, E., Katz, I., Perry, A., & Runyon, E. "Effects of Race of Comparison and Motives on Negro Cognitive Performance." *Journal of Educational Psychology,* 1971, 62, 207-208.

Estes, T.H. & Vaughan, J.L. "Reading Interest and Comprehension: Implications." *Reading Teacher,* 1973, 26, 149-153.

Evans, D.A., Jones, P.K., Wortman, R.A., & Jackson, E.B., Jr. "Traditional Criteria as Predictors of Minority Student Success in Medical School." *Journal of Medical Education,* 1975, *50*, 934-939.

Feitz, R.H. *The MCAT and Success in Medical School.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974. (ERIC Document Reproduction Service, No. ED 090-289).

Flaugher, R.L. *Bias in Testing: A Review and Discussion.* TM Report 36, ERIC Clearinghouse on Tests, Measurement and Evaluation. Princeton, New Jersey: Educational Testing Service, 1974.

Flaugher, R.L. "Project Access Research Report #3: Minority vs. Majority Group Performance on an Aptitude Test Battery." *Educational Testing Service. Research Bulletin RB-71-48,* August, 1971 (entire).

Flaugher, R.L. "The Many Definitions of Test Bias". *American Psychologist,* 1978, *33*, 671-679.

Flaugher, R.L., & Pike, L.W. "Reactions to a Very Difficult Test by an Inner-City High School Population: A Test and Item Analysis." *College Entrance Examination Board Research and Development Reports RDR-69-70, #4, Research Memorandum RM 70-11,* June 1970. Princeton, New Jersey: Educational Testing Service, 1970.

Gordon, T.L., & Johnson, D.G. "Study of U.S. Medical School Applicants, 1975-76." *Journal of Medical Education,* 1977, *52*, 707-730.

Green, R.L., Nyquist, J.G., & Griffore, R.J. *Standardized Achievement Testing: Some Implications for the Lives of Children.* Paper presented at the National Institute of Education Test Bias Conference, Annapolis, Maryland, December 2-5, 1975.

Gross, A.L., & Su, W. "Defining a 'Fair' or 'Unbiased' Selection Model: A Question of Utilities." *Journal of Applied Psychology,* 1975, *60*, 345-351.

Herrnstein, Richard. "I.Q." *Atlantic Monthly,* 1971, *228*, 43-64.

Hersh, J.B. "Effects of Referral Information on Testers." *Journal of Consulting and Clinical Psychology,* 1971 (Aug), *37,* 116-122.

Horst, P. *Psychological Measurement and Prediction.* Belmont, California: Brooks/Cole, 1968.

Jencks, C. et al. *Inequality.* New York: Harper and Row, 1973.

Jensen, A.R. "How Much Can We Boost I.Q. and Scholastic Achievement?" *Harvard Educational Review,* 1969, *39,* 1-123.

Johnston, D.G., & Hutchins, E.B. "Doctor or Dropout? A Study of Medical Student Attrition." *Journal of Medical Education,* 1966, *41,* 1097-1269.

Johnson, D.G., & Sedlacek, W.E. "Retention by Sex and Race of 1968-72 U.S. Medical School Entrants, 1970-72." *Journal of Medical Education,* 1975, *50,* 713-755.

Johnson, L.W. "History of the Education of Negro Physicians." *Journal of Medical Education,* 1967, *42,* 439-446.

Johnson, S.T. "A Statistical and Cognitive Item Analysis of a Test for High-Ability Students." Unpublished Internship Report, Michigan State University Office of Evaluation Services, 1963.

Johnson, S.T. *Issues in Achievement Test Bias - A Response to Robert Ebel.* Statement presented at the National Institute of Education Test Bias Conference, Annapolis, Maryland, December 2-5, 1975.

Katz, I., Atchison, C., Epps, E., & Roberts, S. "Race of Evaluator, Race of Norm, and Expectancy as Determiners of Black Performance." *Journal of Experimental Social Psychology,* 1972, *8,* 1-15.

Katz, I., Roberts, O., & Robinson, J.M. "Effects of Difficulty, Race of Administrator, and Instructions on Negro Digit-Symbol Performance." *ONR Technical Report,* Washington, D.C.: Office of Naval Research, 1963.

Katz, I., Robinson, J.M., Epps, E.G., & Waly, P. "The Influence of Race of the Experimenter and Instructions Upon the Expression of Hostility by Negro Boys." *Journal of Social Issues,* 1964, *20,* 54-59.

Kirkland, M.C. "The Effects of Tests on Students and Schools." *Review of Educational Research,* 1971, *41*, 303-347.

Kotkin, Joel. "Most of the Minority Students Competed Successfully in School," *The Washington Post,* 2 October 1977, p. A12.

Kozol, J. *Death at an Early Age.* Boston: Houghton Mifflin, 1967.

Kruskal, W. "Formulas, Numbers, Words: Statistics in Prose." *The American Scholar,* Spring 1978, *47*, 223-229.

Linn, R.L. "Fair Test Use in Selection." *Review of Educational Research,* 1973, *43*, 139-161.

Litner, A.C., & Ducette, J.D. "The Effects of Locus of Control, Academic Failure, and Task Dimensions on Students Responsiveness to Praise." *American Educational Research Journal,* 1974, *11*, 231-239.

Lloyd, S.M., Johnson, D.G., & Mann, M. "Survey of Graduates of a Traditionally Black College of Medicine." *Journal of Medical Education,* 1978, *53*, 640-650.

Lord, Frederick. "Test Theory and the Public Interest." in *Testing and the Public Interest: Proceedings of the 1976 ETS Invitational Conference.* Princeton, New Jersey: Educational Testing Service, 1977.

McCormack, W. "The Bakke Decision: Implications for Higher Education Admissions." *A Report of the ACE-AALS Committee on Bakke.* Washington, D.C.: American Council on Education and Association of American Law Schools, 1978.

McGuire, K. *Construction and Use of Written Simulations.* New York: Harcourt, Brace & World, 1976.

McMorris, R.F., & Ambrosimo, R.J. "Self-Report Predictors: A Reminder." *Journal of Educational Measurement,* 1973, *10*, 13-18.

Medley, D.M., & Quirk, T.J. "The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teachers Examinations." *Journal of Educational Measurement,* 1974, *11*, 235-245.

Merwin, J.C. "Educational Measurement of What Characteristic of Whom (or What) by Whom or Why." *Journal of Educational Measurement,* 1973, *10*, 1-6.

Miller, L.P. (ed.) *Testing of Black Students: A Symposium.* Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Molidor, J.B., Elstein, A.S., & Krupka, J.W. *The New MCAT: What Does It Measure?* Paper presented at the Annual Meeting of the Association of American Medical Colleges, New Orleans, 1978.

Montoya, R. "Statement to Task Force on Graduate and Professional Admissions, May 12, 1977" in *Task Force on Graduate and Professional Admissions, Final Report. Berkeley: University of California, 1977.*

Novick, M.R. "The Influence of the Law on Professional Measurement Standards" in *Educational Measurement and the Law: Proceedings of the 1977 ETS Invitational Conference.* Princeton, New Jersey: Educational Testing Service, 1978.

Novick, M.R. & Petersen, N.S. "Towards Equalizing Educational and Employment Opportunity." *Journal of Educational Measurement,* 1973, *13*, 77-88.

Petersen, N.S., & Novick, M.R. "An Evaluation of Some Models for Culture-Fair Selection." *Journal of Educational Measurement,* 1976, *13*, 3-30.

Pfeifer, C.M., Jr., & Sedlacek, W.E. "The Validity of Academic Predictors for Black and White Students at a Predominantly White University." *Journal of Educational Measurement,* 1971, *8*, 253-261.

Raup, R.M., & Williams, E.A. "Negro Students in Medical Schools in the United States." *Journal of Medical Education,* 1964, *39*, 444-450.

Reitzes, D.C., & Elkhanialy, H. "Black Students in Medical Schools." *Journal of Medical Education,* 1976, *51*, 1001-1005.

Sawyer, R.L., Cole, N.S., & Cole, J.W.L. "Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection." *Journal of Educational Measurement,* 1973, *13*, 59-76.

Schafer, R. "Psychoanalytic Interpretation in Rorschach Testing: Theory and Application." New York: Grune & Stratton, 1954.

Schofield, W., & Merwin, J.C. "The Use of Scholastic Aptitude, Personality and Interest Test Data in The Selection of Medical Students." *Journal of Medical Education*, 1966, *41*, 502-509.

Sedlacek, W., & Brooks, G. *Racism in American Education: A Model for Change*. Chicago: Nelson-Hall, 1976.

Temp, E. "Validity of the SAT for Blacks and Whites in Thirteen Integrated Institutions." *Journal of Educational Measurement*, 1971, *8*, 245-251.

Thorndike, R.L. "Concepts of Culture Fairness." *Journal of Educational Measurement*, 1971, *8*, 63-70.

Thorndike, R.L. (ed.). *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

U.S. Commission on Civil Rights. *Toward Equal Educational Opportunity: Affirmative Admissions Programs at Law and Medical Schools*. U.S. Commission on Civil Rights Clearinghouse Publication #55, June 1978.

University of California Task Force on Graduate and Professional Admissions (I.M. Heyman, Chairman). *Final Report*. Berkeley: University of California, 1977.

Waldman, Burt. "Economic and Racial Disadvantage as Reflected in Traditional Medical School Selection Factors." *Journal of Medical Education*, 1977, *52*, 961-970.

Walster, E., Cleary, A., & Clifford, M. "The Effects of Race and Sex on College Admission." *Sociology of Education*, 1971, *44*, 237-244.

Weaver, R.C. "The Negro as an American; Address, June 13, 1963." *Vital Speeches*, 1963, *29*, 625-629.

Williams, R.L. "On Black Intelligence." *Journal of Black Studies*, 1973, *4*, 29-39.

Willie, D., Kramer, B., & Brown, B. (eds.). *Racism and Mental Health*. Pittsburgh, Pa.: University of Pittsburgh Press, 1973.

ISEP PUBLICATIONS

**Equal Educational Opportunity for Blacks in U.S. Higher Education: An Assessment*

**The Lengthening Shadow of Slavery: A Historical Justification for Affirmative Action for Blacks in Higher Education,* by Dr. John E. Fleming

Directory of National Sources of Data on Blacks in Higher Education, 1976 Edition

Proceedings from the National Invitational Conference on Race and Ethnic Data, edited by Dr. Elizabeth A. Abramowitz

**The Changing Mood in America: Eroding Commitment?* by Dr. Faustine C. Jones

***Higher Education's Responsibility for Advancing Equality of Opportunity and Justice,* by Dr. James E. Cheek

The Bakke Case Primer

**Equal Educational Opportunity: More Pomise Than Progress*

**The Case for Affirmative Action for Blacks in Higher Education,* by Dr. John E. Fleming, Gerald R. Gill, and Dr. David H. Swinton

Proceedings of the National Invitational Conference on Advancing Equality of Opportunity: A Matter of Justice, edited by Dr. Cynthia J. Smith

Affirmative Action for Blacks in Higher Education: A Report

Forthcoming:-

**Elusive Equality: The Status of Black Americans in Higher Education,* by Dr. Lorenzo Morris

**The Dilemma of Access: Minorities in Two Year Colleges,* by Dr. Michael Olivas

*Available from Howard University Press
**First ISEP occasional paper

For further information, please contact:

Institute for the Study of Educational Policy
Howard University - Dunbarton Campus
2935 Upton St., N.W.
Washington, D.C. 20008
(202) 686-6686